

NECESSARY CONDITIONS FOR OPTIMAL STRATEGIES IN A CLASS OF DIFFERENTIAL GAMES AND CONTROL PROBLEMS*

LEONARD D. BERKOVITZ†

1. Introduction. In [2] we studied a class of differential games having pure strategy solutions. We obtained necessary conditions that must be fulfilled along trajectories resulting from the use of optimal pure strategies and we gave a sufficiency theorem. The necessary conditions were obtained by reducing the problem to two variational problems and then applying the first order necessary conditions for these problems. In the present paper we shall study a slightly wider class of differential games. We shall obtain the necessary conditions of [2] and the properties of the value function W without recourse to the first order necessary conditions for the variational problems. In fact, our arguments will also give the first order necessary conditions for variational problems, or control problems, that have a synthesis of a certain type. Our present methods make use of the assumed existence of optimal strategies, or feedback control laws, instead of the existence of a single optimal trajectory. The developments of §4 of this paper stem from Isaac's "tenet of transition" idea [6]; in the case of control problems this is Bellman's dynamic programming idea [1].

We conclude this introductory section by listing certain definitions and conventions that we shall use throughout the paper. Vector matrix notation will generally be used. Vectors and matrices will be denoted by single letters. Superscripts will be used to denote the components of a vector; subscripts will be used to distinguish vectors. Vectors will be written as matrices consisting of either one row or one column. We shall not use a transpose symbol to distinguish between the two usages, as it will be clear from the context how the vector is to be considered. Thus, the scalar product of two n -dimensional vectors λ and G , say, will be written as λG . All scalars that occur will be real and all vectors will have real components. A vector will be called *positive* if each of its components is positive. Negative, nonpositive and nonnegative vectors are defined similarly.

Let

$$\chi(t, x, y, z) = (\chi^1(t, x, y, z), \dots, \chi^m(t, x, y, z))$$

be a vector-valued function defined and differentiable on a region of

* Received by the editors September 15, 1966, and in revised form November 8, 1966.

† Division of Mathematical Sciences, Purdue University, Lafayette, Indiana 47907. This research was supported by the National Science Foundation under Grant GP-06113.

(t, x, y, z) -space. If z is an s -dimensional vector, we can form a matrix of partial derivatives in which the (i, j) th element is

$$\frac{\partial \chi^i}{\partial z^j}, \quad i = 1, \dots, m, \quad j = 1, \dots, s.$$

The symbol χ_z will denote this matrix. Note that when $m = 1, s = 1$, this reduces to the usual notation for a partial derivative. The symbols χ_x and χ_y will have similar meanings.

The term *region* will mean, as usual, an open connected set. The closure of a region \mathfrak{D} will be denoted by $\bar{\mathfrak{D}}$ and $\bar{\mathfrak{D}}$ will be called a *closed region*. A real-valued function $h(t, x)$ will be said to be of class $C^{(k)}$ in x on \mathfrak{D} if it is continuous in (t, x) in \mathfrak{D} and all of its derivatives with respect to x up to and including those of order k exist on \mathfrak{D} and are continuous in (t, x) on \mathfrak{D} . A real-valued function $h(t, x)$ will be said to be of class $C^{(k)}$ in x on $\bar{\mathfrak{D}}$ if it is $C^{(k)}$ in x on \mathfrak{D} and h and the derivatives of h with respect to x up to and including those of order k have continuous extensions to $\bar{\mathfrak{D}}$. A vector-valued function will be said to be of class $C^{(k)}$ in x on $\bar{\mathfrak{D}}$ if each of its components is $C^{(k)}$ in x on $\bar{\mathfrak{D}}$. A region will be said to have a piecewise smooth boundary if its boundary consists of a finite number of manifolds with boundary. A finite collection of subregions $\mathfrak{D}_1, \dots, \mathfrak{D}_\alpha$ of a region \mathfrak{D} will be said to constitute a *decomposition* of \mathfrak{D} whenever the following conditions hold:

- (i) each $\mathfrak{D}_i, i = 1, \dots, \alpha$, is connected and has a piecewise-smooth boundary;
- (ii) $\mathfrak{D}_i \cap \mathfrak{D}_j = 0$ if $i \neq j$;

A real-valued function defined on $\bar{\mathfrak{D}}$ will be said to be piecewise $C^{(k)}$ in x on $\bar{\mathfrak{D}}$ if there is a decomposition of \mathfrak{D} such that on each \mathfrak{D}_i the function agrees with a function that is $C^{(k)}$ in x on $\bar{\mathfrak{D}}_i$. A vector-valued function will be said to be piecewise $C^{(k)}$ in x on $\bar{\mathfrak{D}}$ if each component is piecewise $C^{(k)}$ in x on $\bar{\mathfrak{D}}$.

The letter t will denote time, and the operator (d/dt) will be denoted by a prime.

2. Formulation of the game. In this section we shall formulate the class of games to which our analysis is applicable. The class of control problems to which the results of this paper are applicable is obtained by suppressing the role of the maximizing player. Analytically this means suppressing y in what follows. We shall leave this to the reader and we shall treat the game situation, as it is the more complicated.

Let x be an n -dimensional vector, let y be an \bar{s} -dimensional vector, and let z be an s -dimensional vector. We shall be concerned with a bounded region \mathfrak{G} of (t, x) -space and a bounded region \mathfrak{S} of (t, x, y, z) -space. We assume that \mathfrak{G} is contained in the projection of \mathfrak{S} into (t, x) -space. We

shall be concerned with a real-valued function $f(t, x, y, z)$ and a vector-valued function

$$G(t, x, y, z) = (G^1(t, x, y, z), \dots, G^n(t, x, y, z)),$$

with range contained in Euclidean n -space. We assume that f and G are of class $C^{(1)}$ on \mathfrak{S} .

We shall also be concerned with a function $\Omega(t, x)$ that maps points (t, x) of \mathfrak{G} into subsets of E^s , and a function $\bar{\Omega}(t, x)$ that maps points of \mathfrak{G} into subsets of \bar{s} -space. We assume that Ω and $\bar{\Omega}$ have the property that $(t, x, \Omega(t, x), \bar{\Omega}(t, x))$ belongs to \mathfrak{S} for all (t, x) in \mathfrak{G} .

Let $\mathfrak{I}_i', i = 1, \dots, \alpha$, be a collection of n -dimensional manifolds of class $C^{(1)}$, each of which lies in \mathfrak{G} and is given parametrically by the equations

$$(2.1) \quad t = T_i(\sigma), \quad x = \chi_i(\sigma),$$

where $\sigma = (\sigma^1, \dots, \sigma^n)$ ranges over a cube in E^n . We select a connected submanifold \mathfrak{I}_i of each \mathfrak{I}_i' , and form

$$\mathfrak{T} = \bigcup_{i=1}^{\alpha} \mathfrak{I}_i.$$

We shall call \mathfrak{T} the *terminal surface* of the game.

Let \mathfrak{G}_1 be a subregion of \mathfrak{G} such that \mathfrak{T} is contained in \mathfrak{G}_1 . Let g be a real-valued function of class $C^{(1)}$ on \mathfrak{G}_1 . The function g is therefore defined on \mathfrak{T} and is of class $C^{(1)}$ on each \mathfrak{I}_i . We shall call g the *terminal payoff* function.

We now define strategies. Let \mathfrak{R} be a region with closure $\bar{\mathfrak{R}}$ contained in \mathfrak{G} , and such that the terminal surface \mathfrak{T} forms a part of the boundary of \mathfrak{R} . Further, assume that \mathfrak{T} separates \mathfrak{G} . The region \mathfrak{R} will be the region of (t, x) -space in which the play of the game takes place. Let \mathfrak{Y} denote the class of functions Y that are piecewise $C^{(1)}$ in x on $\bar{\mathfrak{R}}$, have their range in s -dimensional Euclidean space, and satisfy the condition $Y(t, x) \in \bar{\Omega}(t, x)$ for all $(t, x) \in \bar{\mathfrak{R}}$. Let \mathfrak{Z} denote the class of functions Z that are piecewise $C^{(1)}$ in x on $\bar{\mathfrak{R}}$, have range in E^s , and satisfy $Z(t, x) \in \bar{\Omega}(t, x)$ for all $(t, x) \in \bar{\mathfrak{R}}$. We assume that \mathfrak{Y} and \mathfrak{Z} are nonvoid.

Let $Y \in \mathfrak{Y}$ and $Z \in \mathfrak{Z}$ and consider the differential equation

$$(2.2) \quad \frac{dx}{dt} = G(t, x, Y(t, x), Z(t, x)),$$

subject to the initial condition

$$(2.3) \quad x(\tau) = \xi.$$

If (τ, ξ) is a point of discontinuity of Y or Z , or both Y and Z , there may be more than one solution of (2.2) that satisfies (2.3). Furthermore, a solution that is unique in a neighborhood of τ may bifurcate at some point

of the trajectory at which Y or Z or both are discontinuous. If (τ, ξ) is a point such that both Y and Z are continuous in a neighborhood of (τ, ξ) , then, since G , Y and Z are $C^{(1)}$ functions of x , it follows that there is a unique solution of (2.2) satisfying (2.3) in a neighborhood of τ .

We shall say that $Y \in \mathcal{Y}$ and $z \in \mathcal{Z}$ is a *playable pair* if for every (τ, ξ) in \mathfrak{R} , every solution of (2.2) satisfying (2.3) stays in \mathfrak{R} and reaches \mathfrak{T} in finite time. For each playable pair we can define a possibly multivalued function in R as follows

$$(2.4) \quad P(\tau, \xi, Y, Z) = g(t_f, x_f) + \int_{\tau}^{t_f} f(t, x(t), Y(t, x(t)), Z(t, x(t))) dt,$$

where $x(t)$ is a solution of (2.2) and (2.3) and (t_f, x_f) is the point at which the solution $x(t)$ intersects \mathfrak{T} . We shall call solutions of (2.2) paths. The function P is called the payoff function.

We assume that there exist a nonvoid subclass $\mathcal{Y}_0 \subset \mathcal{Y}$ and a nonvoid subclass $\mathcal{Z}_0 \subset \mathcal{Z}$ such that if $Y \in \mathcal{Y}_0$ and $Z \in \mathcal{Z}_0$, then (Y, Z) is a playable pair. Let $\mathcal{Y}_1, \mathcal{Z}_1$ be the maximal pair of subclasses with this property. We call the functions in \mathcal{Y}_1 and \mathcal{Z}_1 the *pure strategies* for the players.

The game can now be defined. Let (τ, ξ) be given. Player I is to choose a pure strategy Y in \mathcal{Y}_1 so as to maximize $P(\tau, \xi, Y, Z)$; Player II is to choose a pure strategy Z in \mathcal{Z}_1 so as to minimize $P(\tau, \xi, Y, Z)$.

Let (Y^*, Z^*) be a playable pair such that the payoff $P(t, x, Y^*, Z^*)$ is single-valued on \mathfrak{R} . In other words, if there is more than one path starting at a point (t, x) , then all the values $P(t, x, Y^*, Z^*)$ are equal. Denote this common value as $W(t, x)$. Suppose further that, for all Y in \mathcal{Y}_1 and Z in \mathcal{Z}_1 , the inequalities

$$(2.5) \quad P(t, x, Y, Z^*) \leq W(t, x) \leq P(t, x, Y^*, Z)$$

hold for all (t, x) in \mathfrak{R} . We emphasize that this must hold for all values $P(t, x, Y, Z^*)$ and all values $P(t, x, Y^*, Z)$. In this event, we say that (Y^*, Z^*) is a *saddle point* relative to the classes \mathcal{Y}_1 and \mathcal{Z}_1 . The strategies Y^* and Z^* are said to be *optimal*. The function $W(t, x)$ is called the *value* of the game.

The notation $\phi_{(k)}^*(t)$ or $\phi_{(k)}^*(t, \tau, \xi)$ will be used for paths resulting from (Y^*, Z^*) and starting at (τ, ξ) ; such paths will be called *optimal paths*. The subscript is used to distinguish the different paths that may start at (τ, ξ) .

In this paper we shall assume that the game has a saddle point (Y^*, Z^*) . We shall impose further restrictions on (Y^*, Z^*) and shall deduce various properties of the solution for this class of games. In order to state the conditions we introduce the notion of a *regular decomposition* of \mathfrak{R} . This is the same definition used in [2]; reference to Fig. 1 will prove helpful in understanding what is meant.

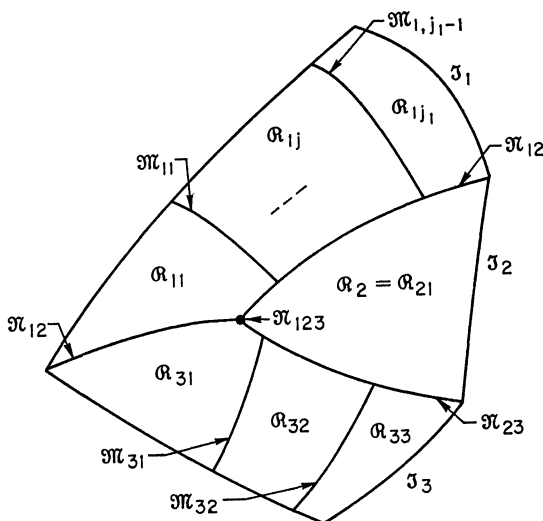


FIG. 1

We define a *regular decomposition* of a region \mathfrak{R} to be a decomposition in which the constituent subregions can be designated

$$\mathfrak{R}_{11}, \mathfrak{R}_{12}, \dots, \mathfrak{R}_{1j_1}, \mathfrak{R}_{21}, \dots, \mathfrak{R}_{2j_2}, \dots, \mathfrak{R}_{\alpha 1}, \dots, \mathfrak{R}_{\alpha j_\alpha},$$

in such a way (see Fig. 1) that the following conditions are satisfied:

(i) The regions \mathfrak{R}_i , defined for each $i = 1, \dots, \alpha$ by the formula

$$\mathfrak{R}_i = \bigcup_{j=1}^{j_i} \mathfrak{M}_{ij} \cap \mathfrak{R},$$

constitute a decomposition of \mathfrak{R} .

(ii) For each $i = 1, \dots, \alpha$, \mathfrak{R}_i always lies on the same side of \mathfrak{T}_i , and $\mathfrak{R}_i \cap \mathfrak{T}_k = \emptyset$ whenever $i \neq k$.

(iii) For each $i = 1, \dots, \alpha$, $\mathfrak{M}_{ij_i} \cap \mathfrak{T}_i \neq \emptyset$, $\mathfrak{M}_{ij} \cap \mathfrak{T}_i = \emptyset$ whenever $j \neq j_i$.

(iv) For each $i = 1, \dots, \alpha$ and $j = 1, \dots, j_i - 1$, the set

$$\mathfrak{M}_{ij} = (\mathfrak{M}_{ij} \cap \mathfrak{M}_{i,j+1}) \cap \mathfrak{R}_i$$

is a connected and oriented manifold of dimension n and class $C^{(1)}$. We suppose that \mathfrak{M}_{ij} can be described by equations

$$(2.6) \quad t = T_{ij}(\sigma), \quad x = \chi_{ij}(\sigma),$$

where $\sigma = (\sigma^1, \dots, \sigma^n)$ ranges over a cube \mathfrak{R}_{ij} in Euclidean n -space.

(v) Each manifold \mathfrak{M}_{ij} divides \mathfrak{R}_i into two disjoint regions such that

each region lies entirely on one side of \mathfrak{M}_{ij} . Furthermore, $\overline{\mathfrak{M}}_{ij} \cap \overline{\mathfrak{M}}_{ik} = 0$ for $i = 1, \dots, \alpha$, and $j, k = 1, \dots, j_i - 1$.

(vi) For each subset i_1, \dots, i_k of the integers $1, \dots, \alpha$, the set $\mathfrak{N}_{i_1, \dots, i_k}$, defined by the formula

$$\mathfrak{N}_{i_1, \dots, i_k} = (\overline{\mathfrak{N}}_{i_1} \cap \overline{\mathfrak{N}}_{i_2} \cap \dots \cap \overline{\mathfrak{N}}_{i_k}) \cap \mathfrak{R},$$

is either the null set or a connected, nonsingular oriented differentiable manifold.

We remark here that, as in the definition of \mathfrak{T}_i , we assume that all of \mathfrak{M}_{ij} can be coordinatized by parameters in a single \mathfrak{R}_{ij} . This is made for the purpose of simplifying the exposition.

We now introduce additional notation that will be used in the sequel. In order to simplify the exposition, we shall sometimes denote the terminal manifolds \mathfrak{T}_i as \mathfrak{M}_{ij_i} . Also, we define \mathfrak{M}_{i0} to be the union of all manifolds $\mathfrak{N}_{i_1, \dots, i_k}$ such that i belongs to the set i_1, \dots, i_k . We define

$$\begin{aligned} \mathfrak{N}_{ij}^+ &= \mathfrak{R}_{ij} \cup \mathfrak{M}_{ij}, \\ \mathfrak{N}_{ij}^- &= \mathfrak{R}_{ij} \cup \mathfrak{M}_{i, j-1}, \\ \tilde{\mathfrak{N}}_{ij} &= \mathfrak{R}_{ij} \cup \mathfrak{M}_{i, j-1} \cup \mathfrak{M}_{ij}, \end{aligned} \tag{2.7}$$

where $i = 1, \dots, \alpha$, and $j = 1, \dots, j_i$.

We now formulate the restriction on (Y^*, Z^*) . Since Y^* and Z^* are piecewise $C^{(1)}$ in x there is associated with this pair of functions a decomposition of \mathfrak{R} , say $\tilde{\mathfrak{R}} = \bigcup \mathfrak{D}_k$, such that in the closure of each \mathfrak{D}_k , Y^* and Z^* are $C^{(1)}$ in x . The saddle restriction is the following:

(i) The decomposition associated with (Y^*, Z^*) is a regular decomposition.

(ii) If (τ, ξ) is a point of \mathfrak{N}_{ij}^- , $i = 1, \dots, \alpha$, $j = 1, \dots, j_i$, then there is a unique optimal path

$$\phi^*(t, \tau, \xi) \equiv \phi_{(i)}^*(t, \tau, \xi) \quad \text{in } \mathfrak{R}_i \quad \text{for } \tau < t < t_{i, j_i}, \tag{2.8}$$

where t_{i, j_i} is the time at which the path reaches \mathfrak{T}_i . Moreover, the path is never tangent to a manifold \mathfrak{M}_{ik} , $k = j, \dots, j_i$, or to a manifold $\mathfrak{N}_{i_1, \dots, i_r}$.

Remarks. If (τ, ξ) is a point of a manifold $\mathfrak{N}_{i_1, \dots, i_r}$, then (τ, ξ) can belong to regions \mathfrak{N}_{ij}^- corresponding to several values of i . Condition (ii) of (2.8) holds for each of these values of i .

In general, at points t_{ik} at which the trajectory intersects a manifold \mathfrak{M}_{ik} , $k = j, \dots, j_{i-1}$, the derivative $\phi^{*'} will be discontinuous, since the manifolds \mathfrak{M}_{ik} are manifolds of discontinuity of at least one of the functions Y^* or Z^* . Condition (ii) is to be understood in the sense that neither $(1, \phi^{*'}(t_{ik} + 0, \tau, \xi))$ nor $(1, \phi^{*'}(t_{ik} - 0, \tau, \xi))$ is a tangent vector to \mathfrak{M}_{ik} .$

3. Properties of $\phi^*(t, \tau, \xi)$. Consider a subregion \mathfrak{R}_i . Let (τ, ξ) be a point of \mathfrak{R}_{ij} for some $1 \leq j \leq j_i$. Then there is precisely one optimal path in \mathfrak{R}_i starting at (τ, ξ) and terminating at $\mathfrak{T}_i \equiv \mathfrak{M}_{ij_i}$ that results from (Y^*, Z^*) . We drop the subscript i and denote this path by $\phi^*(t, \tau, \xi)$. We shall study the dependence of ϕ^* on ξ .

For each $j = 1, \dots, j_i$, Y^* and Z^* are $C^{(1)}$ in x on \mathfrak{R}_{ij} and each \mathfrak{M}_{ij} and each $\mathfrak{M}_{i_1, \dots, i_k}$ is $C^{(1)}$. It therefore follows that for each $j = 1, \dots, j_i$ there exist functions Y_{ij}^* and Z_{ij}^* and a region \mathfrak{R}_{ije} containing \mathfrak{R}_{ij} in its interior such that Y_{ij}^* and Z_{ij}^* are $C^{(1)}$ in x on \mathfrak{R}_{ije} and such that for (t, x) in \mathfrak{R}_{ij}

$$Y^*(t, x) = Y_{ij}^*(t, x), \quad Z^*(t, x) = Z_{ij}^*(t, x).$$

Let us fix j and consider the following differential equation on \mathfrak{R}_{ije} :

$$(3.1) \quad \frac{dx}{dt} = G(t, x, Y_{ij}^*(t, x), Z_{ij}^*(t, x)), \quad x(\tau) = \xi,$$

where (τ, ξ) is in \mathfrak{R}_{ije} . Since G is $C^{(1)}$ in (x, y, z) , it follows from the properties of $Y_{ij}^*(t, x)$ and $Z_{ij}^*(t, x)$ and from standard theorems in the theory of ordinary differential equations that there is a function $\psi_j(t, \tau, \xi)$ defined for $a(\tau, \xi) < t < b(\tau, \xi)$ such that ψ_j is the unique solution of (3.1). Moreover, for (τ, ξ) in \mathfrak{R}_{ije} and $a(\tau, \xi) < t < b(\tau, \xi)$ the functions $\psi_{j\xi}(t, \tau, \xi)$, $\psi_{j\tau}(t, \tau, \xi)$, $\psi'_{j\xi}(t, \tau, \xi)$ and $\psi'_{j\tau}(t, \tau, \xi)$ exist and are continuous. (See, for example, [4, Theorem 12, p. 165]).

On \mathfrak{R}_{ij} the right-hand side of (3.1) agrees with the right-hand side of

$$(3.2) \quad \frac{dx}{dt} = G(t, x, Y^*(t, x), Z^*(t, x)), \quad x(\tau) = \xi.$$

Therefore, from our assumption that there is an optimal path from each point in \mathfrak{R}_{ij} and from the uniqueness theorem for differential equations it follows that for each (τ, ξ) in \mathfrak{R}_{ij} there exists a $\delta = \delta(\tau, \xi) > 0$ such that the solutions $\psi_j(t, \tau, \xi)$ of (3.1) are defined and are unique on an interval $(t_{i,j-1} - \delta, t_{i,j} + \delta)$, where t_{ij} is the value of t at which the graph of ψ_j intersects \mathfrak{M}_{ij} , and $t_{i,j-1}$ is the value of t at which the graph of ψ_j intersects $\mathfrak{M}_{i,j-1}$.

For (τ, ξ) in \mathfrak{R}_{ij} the function $\phi^*(t, \tau, \xi)$ is defined for $\tau \leq t \leq t_i(\tau, \xi)$, where t_i is the time at which the trajectory intersects \mathfrak{T} . Moreover, ϕ^* satisfies the differential equation (3.2). Since the right-hand sides of (3.1) and (3.2) agree on \mathfrak{R}_{ij} , it follows from the uniqueness of solutions of (3.1) that for $\tau \leq t \leq t_{ij}$, we have $\phi^*(t, \tau, \xi) = \psi_j(t, \tau, \xi)$. For $t_{i,j-1} \leq t \leq \tau$, we can define $\phi^*(t, \tau, \xi) = \psi_j(t, \tau, \xi)$. Thus, for (τ, ξ) in \mathfrak{R}_{ij} and $t_{i,j-1} \leq t \leq t_{i,j}$, we look upon $\phi^*(t, \tau, \xi)$ as the restriction of $\psi_j(t, \tau, \xi)$ to this proper subdomain of definition. One should keep this

point of view in mind in some of the subsequent developments. Note that ϕ^* is the optimal path through every point on its graph.

The following lemma is a consequence of the statements in the last three paragraphs.

LEMMA 1. *For each (τ, ξ) in $\tilde{\mathfrak{N}}_{ij}$ there are a $t_{i,j-1} = t_{i,j-1}(\tau, \xi)$ and a $t_{i,j} = t_{i,j}(\tau, \xi)$ such that for $t_{i,j-1} \leq t \leq t_{i,j}$, the function $\phi^*(t, \tau, \xi)$ is defined and satisfies (3.2). Moreover, for (τ, ξ) in \mathfrak{N}_{ij} and $t_{i,j-1} \leq t \leq t_{i,j}$, the functions $\phi_\tau^*(t, \tau, \xi)$, $\phi_\xi^*(t, \tau, \xi)$, $\phi_\xi^{*'}(t, \tau, \xi)$ and $\phi_\tau^{*'}(t, \tau, \xi)$ exist and are continuous.*

Remark. If (τ, ξ) belongs to either $\mathfrak{M}_{i,j-1}$ or \mathfrak{M}_{ij} , or if $t = t_{i,j-1}$ or $t = t_{i,j}$, then the values of the above functions can either be thought of as limits computed from the interior, or as values of the appropriate derivatives of $\psi_j(t, \tau, \xi)$.

Let (τ, ξ) be a point of \mathfrak{N}_{ij} , where $1 \leq j \leq j_i$. Then the optimal path $\phi^*(t, \tau, \xi)$ will intersect one of the manifolds \mathfrak{M}_{ij} at a point with coordinates (t_{ij}, x_{ij}) , where

$$x_{ij} = x_{ij}(\tau, \xi) = \phi^*(t_{ij}(\tau, \xi), \tau, \xi).$$

From (2.6) we have that the value of the parameter σ corresponding to this point of \mathfrak{M}_{ij} must satisfy

$$(3.3) \quad t_{ij}(\tau, \xi) = T_{ij}(\sigma), \quad x_{ij}(\tau, \xi) = \chi_{ij}(\sigma),$$

and so must satisfy

$$(3.4) \quad \phi^*(T_{ij}(\sigma), \tau, \xi) - \chi_{ij}(\sigma) = 0.$$

LEMMA 2. *Equation (3.4) defines σ as a $C^{(1)}$ function of (τ, ξ) for (τ, ξ) in \mathfrak{N}_{ij} . The coordinates (t_{ij}, x_{ij}) are $C^{(1)}$ functions of (τ, ξ) .*

We first note that since T_{ij} and χ_{ij} are $C^{(1)}$ functions of σ , it follows that the second statement is a consequence of (3.3) and the first statement. We now prove the first statement.

Equation (3.4) is equivalent to

$$(3.5) \quad \psi_j(T_{ij}(\sigma), \tau, \xi) - \chi_{ij}(\sigma) = 0.$$

Therefore, to prove the first statement we need only show that the Jacobian matrix of (3.5) with respect to σ is nonsingular. The conclusion will then follow from the implicit function theorem. The Jacobian matrix of (3.5) is

$$\begin{aligned} M &= \psi_j'(t_{ij}, \tau, \xi) \frac{\partial T_{ij}}{\partial \sigma} - \frac{\partial \chi_{ij}}{\partial \sigma} \\ &= \phi^{*'}(t_{ij}, 0, \tau, \xi) \frac{\partial T_{ij}}{\partial \sigma} - \frac{\partial \chi_{ij}}{\partial \sigma}. \end{aligned}$$

The matrix

$$N = \begin{pmatrix} 1 & \phi^{*'}(t_{ij} - 0, \tau, \xi) \\ -\frac{\partial T_{ij}}{\partial \sigma} & -\frac{\partial \chi_{ij}}{\partial \sigma} \end{pmatrix}$$

has rank $n + 1$ by virtue of the assumption that the optimal path is not tangent to the manifold \mathfrak{M}_{ij} . Since we can transform N into the matrix

$$\begin{pmatrix} 1 & \phi^{*'}(t_{ij} - 0, \tau, \xi) \\ 0 & M \end{pmatrix}$$

by means of elementary row transformations, it follows that the Jacobian matrix M has rank n , and the lemma is proved.

Remark. The use of ψ_j in the proof shows that we can consider t_{ij} and x_{ij} to be $C^{(1)}$ functions of (τ, ξ) on \mathfrak{R}_{ij} , where if (τ, ξ) is a point of $\mathfrak{M}_{i,j-1}$ or \mathfrak{M}_{ij} , then we assign to $t_{ij}(\tau, \xi)$ and $x_{ij}(\tau, \xi)$ the limits of t_{ij} and x_{ij} as we approach (τ, ξ) from the interior.

Let (τ, ξ) again be a point of \mathfrak{R}_{ij} , where we now suppose that $j < j_i$. We consider the optimal path $\phi^*(t, \tau, \xi)$ in the region $\mathfrak{R}_{i,j+1}$. Denote the coordinates of the point at which the path intersects $\mathfrak{M}_{i,j+1}$ by $(t_{i,j+1}(\tau, \xi), x_{i,j+1}(\tau, \xi))$. Henceforth, we shall not always indicate the dependence of $t_{i,j+1}$ and $x_{i,j+1}$ on (τ, ξ) . Consider $\phi^*(t, \tau, \xi)$ on the interval $t_{ij} \leq t \leq t_{i,j+1}$. On this interval we have

$$(3.6) \quad \phi^*(t, \tau, \xi) = \phi^*(t, t_{ij}, x_{ij}).$$

As in the case of the region \mathfrak{R}_{ij} , there exists a region $\mathfrak{R}_{i,j+1,e}$ containing $\mathfrak{R}_{i,j+1}$ in its interior, and there exist functions Y_{j+1}^* and Z_{j+1}^* that are $C^{(1)}$ in x on $\mathfrak{R}_{i,j+1,e}$ such that $Y_{j+1}^* = Y^*$ and $Z_{j+1}^* = Z^*$ on $\mathfrak{R}_{i,j+1}$. For (α, β) in $\mathfrak{R}_{i,j+1,e}$, let $\psi_{j+1}(t, \alpha, \beta)$ be the solution of

$$\frac{dx}{dt} = G(t, x, Y_{j+1}^*(t, x), Z_{j+1}^*(t, x)), \quad x(\alpha) = \beta.$$

The functions ψ_{j+1} and ψ'_{j+1} possess partial derivatives with respect to α and β that are continuous with respect to (t, α, β) on the domain of definition of ψ . By the arguments used in Lemma 1, we have that for $t_{ij} \leq t \leq t_{i,j+1}$,

$$\phi^*(t, t_{ij}, x_{ij}) = \psi_{j+1}(t, t_{ij}, x_{ij}).$$

Hence, by (3.6),

$$(3.7) \quad \phi^*(t, \tau, \xi) = \psi_{j+1}(t, t_{ij}, x_{ij}), \quad t_{ij} \leq t \leq t_{i,j+1}.$$

From the differentiability properties of ψ_{j+1} , from (3.7) and from Lemma 2, it follows that for (τ, ξ) in \mathfrak{R}_{ij} and $t_{ij} \leq t \leq t_{i,j+1}$, the functions ϕ_τ^* , ϕ_ξ^* , $\phi_\tau^{*'}$ and $\phi_\xi^{*'}$ exist and are continuous functions of (t, τ, ξ) . At $t = t_{ij}$

and $t = t_{i,j+1}$ the values of these functions are taken as limits from the interior of $\mathfrak{R}_{i,j+1}$. We also note that we can consider these functions to be defined and continuous for (τ, ξ) in \mathfrak{N}_{ij} by assigning appropriate limits as function values if (τ, ξ) belongs to $\mathfrak{M}_{i,j-1}$ or \mathfrak{M}_{ij} . Furthermore,

$$\phi_{\xi}^{*}(t, \tau, \xi) = \psi_{j+1,\alpha} \frac{\partial t_{ij}}{\partial \xi} + \psi_{j+1,\beta} \frac{\partial x_{ij}}{\partial \xi},$$

where $\psi_{j+1,\alpha}$ and $\psi_{j+1,\beta}$ are evaluated at $(t, \alpha, \beta) = (t, t_{ij}, x_{ij})$ and $\partial t_{ij}/\partial \xi$ and $\partial x_{ij}/\partial \xi$ are evaluated at (τ, ξ) . In particular, at $t = t_{ij}$ we have

$$(3.8) \quad \phi_{\xi}^{*}(t_{ij} + 0, \tau, \xi) = \psi_{j+1,\alpha}(t_{ij}, t_{ij}, x_{ij}) \frac{\partial t_{ij}}{\partial \xi} + \psi_{j+1,\beta}(t_{ij}, t_{ij}, x_{ij}) \frac{\partial x_{ij}}{\partial \xi}.$$

From standard theorems in the theory of ordinary differential equations' we have

$$\psi_{j+1,\alpha}(t_{ij}, t_{ij}, x_{ij}) = -G(P_{ij} + 0),$$

$$\psi_{j+1,\beta}(t_{ij}, t_{ij}, x_{ij}) = I,$$

where we have set

$$(3.9) \quad P_{ij} + 0 = (t_{ij}, x_{ij}, Y_{i,j+1}^{*}(t_{ij}, x_{ij}), Z_{i,j+1}^{*}(t_{ij}, x_{ij}))$$

and I is the identity matrix. Substituting these quantities back into (3.8), we get the relation

$$(3.10) \quad \phi_{\xi}^{*}(t_{ij} + 0, \tau, \xi) = -G(P_{ij} + 0) \frac{\partial t_{ij}}{\partial \xi} + \frac{\partial x_{ij}}{\partial \xi}.$$

We shall now compute $\phi_{\xi}^{*}(t_{ij} - 0, \tau, \xi)$. If we substitute $\sigma = \sigma(\tau, \xi)$ into (3.5) we obtain an identity in (τ, ξ) . If we differentiate this identity with respect to ξ and use (3.3), we get

$$\psi_j'(t_{ij}, \tau, \xi) \frac{\partial t_{ij}}{\partial \xi} + \psi_{j\xi} - \frac{\partial x_{ij}}{\partial \xi} = 0.$$

Since for (τ, ξ) in \mathfrak{N}_{ij} and $t_{i,j-1} \leq t \leq t_{i,j}$ we have $\phi^{*}(t, \tau, \xi) = \psi_j(t, \tau, \xi)$, we can rewrite the last equation as follows:

$$\phi_{\xi}^{*}(t_{ij} - 0, \tau, \xi) = -\phi^{*'}(t_{ij} - 0, \tau, \xi) \frac{\partial t_{ij}}{\partial \xi} + \frac{\partial x_{ij}}{\partial \xi}.$$

If we set

$$(3.11) \quad P_{ij} - 0 = (t_{ij}, x_{ij}, Y_{ij}^{*}(t_{ij}, x_{ij}), Z_{ij}^{*}(t_{ij}, x_{ij})),$$

and use the fact that ϕ^* satisfies (3.1) in \mathfrak{R}_{ij} , we finally obtain

$$(3.12) \quad \phi_{\xi}^*(t_{ij} - 0, \tau, \xi) = -G(P_{ij} - 0) \frac{\partial t_{ij}}{\partial \xi} + \frac{\partial x_{ij}}{\partial \xi}.$$

Comparing the last relation with (3.10) we obtain

$$(3.13) \quad \begin{aligned} \phi_{\xi}^*(t_{ij} + 0, \tau, \xi) - \phi_{\xi}^*(t_{ij} - 0, \tau, \xi) \\ = -[G(P_{ij} + 0) - G(P_{ij} - 0)] \frac{\partial t_{ij}}{\partial \xi}. \end{aligned}$$

We can use the above procedures to establish the properties of ϕ^* in each of the regions \mathfrak{R}_{ik} , $j \leq k \leq j_i$, and to determine the one-sided limits of ϕ_{ξ}^* at the points $(t_{ik}, x_{ik}) = (t_{ik}(\tau, \xi), x_{ik}(\tau, \xi))$ of intersection with the manifolds \mathfrak{M}_{ik} . In the course of this procedure we establish the results of Lemma 2 at each of the manifolds \mathfrak{M}_{ik} . In particular, it follows that the matrix M with j replaced by k has rank n . We summarize the results of this procedure in the following lemma.

LEMMA 3. *Let (τ, ξ) belong to \mathfrak{R}_{ij} , $1 \leq j \leq j_i$, and let $(t_{ik}, x_{ik}) = (t_{ik}(\tau, \xi), x_{ik}(\tau, \xi))$ denote the point at which ϕ^* intersects the manifold \mathfrak{M}_{ik} , $j \leq k \leq j_i$. Then for each $j \leq k \leq j_i$, the functions $\phi^*(t, \tau, \xi)$, $\phi_{\xi}^*(t, \tau, \xi)$, $\phi_{\tau}^*(t, \tau, \xi)$, $\phi_{\tau}^{*'}(t, \tau, \xi)$, $\phi_{\xi}^{*'}(t, \tau, \xi)$ exist and are continuous for (τ, ξ) in \mathfrak{R}_{ij} and $t_{i,k-1} \leq t \leq t_{i,k}$, where at $t = t_{i,k-1}$ and $t = t_{i,k}$ the values assigned to the functions are their limits from the interior. Moreover, if at points $(\hat{\tau}, \hat{\xi})$ belonging to $\mathfrak{M}_{i,j-1}$ or \mathfrak{M}_{ij} we assign to these functions the limits as $(\tau, \xi) \rightarrow (\hat{\tau}, \hat{\xi})$ from the interior of \mathfrak{R}_{ij} , then the resulting functions are well defined and continuous for $(\tau, \xi) \in \mathfrak{R}_{ij}$. At $t = t_{ik}$ the one-sided limits of ϕ_{ξ}^* are given by (3.10) and (3.12) with j replaced by k .*

For each $k = j, j+1, \dots, j_i$, the functions $t_{ik}(\tau, \xi)$ and $x_{ik}(\tau, \xi)$ as well as the parameter values σ corresponding to these points of the manifold \mathfrak{M}_{ik} are $C^{(1)}$ functions of (τ, ξ) in \mathfrak{R}_{ij} . The matrices

$$N_k = \phi^{*'}(t_{ik} - 0, \tau, \xi) \frac{\partial T_{ik}}{\partial \sigma} - \frac{\partial \chi_{ik}}{\partial \sigma}, \quad k = j, \dots, j_i,$$

are nonsingular. Also

$$(3.14) \quad \frac{\partial t_{ik}}{\partial \xi} = \frac{\partial T_{ik}}{\partial \sigma} \frac{\partial \sigma}{\partial \xi}, \quad \frac{\partial x_{ik}}{\partial \xi} = \frac{\partial \chi_{ik}}{\partial \sigma} \frac{\partial \sigma}{\partial \xi}.$$

4. The value function. Let (τ, ξ) be a point of \mathfrak{R}_{ij} , $1 \leq j \leq j_i$. Then

$$(4.1) \quad W(\tau, \xi) = g + \int_{\tau}^{t_{ij}} \bar{f} dt + \sum_{k=j}^{j_i-1} \int_{t_{ik}}^{t_{i,k+1}} \bar{f} dt,$$

where g is evaluated at (t_{ij}, x_{ij}) , the endpoint of the optimal path, and

$$(4.2) \quad \bar{f} = f(t, \phi^*(t, \tau, \xi), Y^*(t, \phi^*(t, \tau, \xi)), Z^*(t, \phi^*(t, \tau, \xi))).$$

From these formulas and from Lemma 3 it follows that W_τ and W_ξ exist and are continuous on R_{ij} . It also follows that if $(\tau, \xi) \rightarrow (\tau_0, \xi_0)$, where (τ_0, ξ_0) belongs to $\mathcal{M}_{i,j-1}$ or \mathcal{M}_{ij} , then

$$\lim W_\xi(\tau, \xi) \quad \text{and} \quad \lim W_\tau(\tau, \xi)$$

both exist and the functions that we get by assigning the values of these limits at such points (τ_0, ξ_0) are continuous on \mathfrak{R}_{ij} .

Our next objective is to establish a partial differential equation that W satisfies on R_{ij} . This equation was first obtained heuristically by Isaacs [5]. A more readily available version of his work is to be found in his book [6].

Let (τ, ξ) be a point of \mathfrak{R}_{ij} and let Z be any strategy in \mathfrak{Z}_1 . Then it is not hard to see that there exists a closed region $N(\tau, \xi)$ with the following properties:

- (i) $N(\tau, \xi) \subset \mathfrak{R}_{ij}$;
- (ii) $(\tau, \xi) \in N(\tau, \xi)$;
- (iii) there is a $\gamma > 0$ such that $\tau \leq t \leq \tau + \gamma$ is contained in the projection of $N(\tau, \xi)$ on the t -axis;
- (iv) the function

$$\hat{Z}(t, x) = \begin{cases} Z^*(t, x) & \text{if } (t, x) \notin N(\tau, \xi), \\ Z(t, x) & \text{if } (t, x) \in N(\tau, \xi) \end{cases}$$

is in \mathfrak{Z}_1 .

Let $\psi(t)$ denote a path resulting from (Y^*, \hat{Z}) and starting at (τ, ξ) . Let $\tau + \delta$ be the last time at which this path leaves $N(\tau, \xi)$. Then from the definition of \hat{Z} it follows that for $t \geq \tau + \delta$, $\psi(t) = \phi^*(t, \tau + \delta, x(\tau + \delta))$.

From the hypothesis that (Y^*, Z^*) is a saddle point and from the above remarks we have that

$$\begin{aligned} W(\tau, \xi) &= P(\tau, \xi, Y^*, Z^*) \leq P(\tau, \xi, Y^*, \hat{Z}) \\ &= g(t_f, x_f) + \left(\int_\tau^{\tau+\delta} + \int_{\tau+\delta}^{t_f} \right) f(t, \psi(t), Y^*(t, \psi(t)), \hat{Z}(t, \psi(t))) dt \\ &= \int_\tau^{\tau+\delta} f(t, \psi(t), Y^*(t, \psi(t)), \hat{Z}(t, \psi(t))) dt + \\ &\quad W(\tau + \delta, \psi(\tau + \delta)). \end{aligned}$$

Hence

$$(4.3) \quad -W(\tau + \delta, \psi(\tau + \delta)) + W(\tau, \xi) \leq \int_\tau^{\tau+\delta} f(t, \psi, Y^*(t, \psi), \hat{Z}(t, \psi)) dt.$$

We note that if $Z = Z^*$, then $\hat{Z} = Z^*$ and the equality holds in (4.3).

We shall now let $N(\tau, \xi) \rightarrow (\tau, \xi)$ by letting diameter $N(\tau, \xi) \rightarrow 0$; this clearly implies that $\delta \rightarrow 0$. Since W_τ and W_ξ are continuous on \mathfrak{R}_{ij} and (τ, ξ) is an interior point of \mathfrak{R}_{ij} , we can apply the mean value theorem to the left-hand side of (4.3) and rewrite it as follows:

$$-W_\tau(\tau, \xi)\delta - W_\xi(\tau, \xi)(\psi(\tau + \delta) - \xi) + o(1)\delta + o(1)(\psi(\tau + \delta) - \xi).$$

Since

$$\begin{aligned}\psi(\tau + \delta) - \xi &= \psi(\tau + \delta) - \psi(\tau) \\ &= (\psi'(\tau) + o(1))\delta \\ &= (G(\tau, \xi, Y^*(\tau, \xi), \hat{Z}(\tau, \xi)) + o(1))\delta,\end{aligned}$$

and since $\hat{Z}(\tau, \xi) = Z(\tau, \xi)$, we can further rewrite the left-hand side of (4.3) in the form

$$-(W_\tau(\tau, \xi) + W_\xi(\tau, \xi)G(\tau, \xi, Y^*(\tau, \xi), Z(\tau, \xi)))\delta + o(\delta).$$

The right-hand side of (4.3) can be written in the form

$$(f(\tau, \xi, Y^*(\tau, \xi), Z(\tau, \xi)) + o(1))\delta.$$

If, after making these transformations, we divide (4.3) by $\delta > 0$ and then let $N(\tau, \xi) \rightarrow (\tau, \xi)$, we obtain

$$(4.4) \quad -W_\tau(\tau, \xi) \leq f(\tau, \xi, Y^*(\tau, \xi), z) + W_\xi(\tau, \xi)G(\tau, \xi, Y^*(\tau, \xi), z),$$

where we have put $z = Z(\tau, \xi)$. Since Z is an arbitrary element of \mathfrak{Z}_1 , we have the result that (4.4) holds for all z such that $z = Z(\tau, \xi)$ for some Z in \mathfrak{Z}_1 . Since we have equality holding in (4.3) for $Z = Z^*$, we have

$$(4.5) \quad -W_\tau(\tau, \xi) = f(\tau, \xi, y^*, z^*) + W_\xi(\tau, \xi)G(\tau, \xi, y^*, z^*),$$

where we have put $z^* = Z^*(\tau, \xi)$.

If we now consider strategies Y in \mathfrak{Y}_1 against Z^* and carry out arguments similar to the above, we arrive at the following inequality instead of (4.4):

$$(4.6) \quad -W_\tau(\tau, \xi) \geq f(\tau, \xi, y, Z^*(\tau, \xi)) + W_\xi(\tau, \xi)G(\tau, \xi, y, Z^*(\tau, \xi)),$$

for all y such that $y = Y(\tau, \xi)$ for some Y in \mathfrak{Y}_1 .

Combining (4.4), (4.5), and (4.6), we get

$$\begin{aligned}(4.7) \quad & f(\tau, \xi, z^*, y) + W_\xi(\tau, \xi)G(\tau, \xi, z^*, y) \\ & \leq f(\tau, \xi, z^*, y^*) + W_\xi(\tau, \xi)G(\tau, \xi, y^*, z^*) \\ & = -W_\tau(\tau, \xi) \leq f(\tau, \xi, z, y^*) + W_\xi(\tau, \xi)G(\tau, \xi, z, y^*).\end{aligned}$$

As is well known, this implies the relation

$$\begin{aligned}
 \max_y \min_z (f + W_\xi G) &= \min_z \max_y (f + W_\xi G) \\
 (4.8) \qquad \qquad \qquad &= f(\tau, \xi, y^*, z^*) + W_\xi G(\tau, \xi, y^*, z^*) \\
 &= -W_\tau(\tau, \xi),
 \end{aligned}$$

where

$$\begin{aligned}
 y &\in E[y \mid y = Y(\tau, \xi), Y \in \mathfrak{Y}_1], \\
 z &\in E[z \mid z = Z(\tau, \xi), Z \in \mathfrak{Z}_1].
 \end{aligned}$$

Relation (4.8) and its extensions constitute one of the principal results of this section.

An equivalent way of stating (4.8) is the following. At each point (τ, ξ) of \mathfrak{R}_{ij} consider the game $\Gamma(\tau, \xi)$ with payoff

$$f(\tau, \xi, y, z) + W_\xi(\tau, \xi)G(\tau, \xi, y, z),$$

where the set of strategies for Player I consists of all vectors y such that $y = Y(\tau, \xi)$ for some $Y \in \mathfrak{Y}_1$ and the set of strategies for Player II consists of all vectors z such that $z = Z(\tau, \xi)$ for some $Z \in \mathfrak{Z}_1$. Then $\Gamma(\tau, \xi)$ has a pure strategy solution $(y^*, z^*) = (Y^*(\tau, \xi), Z^*(\tau, \xi))$. Moreover, the value of the game is equal to $-W_\tau(\tau, \xi)$.

Equations (4.7) and (4.8) were derived under the assumption that (τ, ξ) is an interior point of a region \mathfrak{R}_{ij} . It therefore follows by continuity that if (τ, ξ) is a point of a manifold \mathfrak{M}_{ij} , then (4.7) and (4.8) hold with $(y^*, z^*) = (Y_{ij}^*(\tau, \xi), Z_{ij}^*(\tau, \xi))$ and with

$$\begin{aligned}
 W_\tau(\tau, \xi) &= W_\tau^-(\tau, \xi) = \lim W_\tau(t, x), \\
 W_\xi(\tau, \xi) &= W_\xi^-(\tau, \xi) = \lim W_\xi(t, x),
 \end{aligned}$$

where the limit is taken as $(t, x) \rightarrow (\tau, \xi)$ from \mathfrak{R}_{ij} . If $j < j_i$, it also follows that (4.7) and (4.8) hold with $(y^*, z^*) = (Y_{i, j+1}^*(\tau, \xi), Z_{i, j+1}^*(\tau, \xi))$ and with $W_\tau = W_\tau^+$ and $W_\xi = W_\xi^+$, where W_τ^+ and W_ξ^+ are limits from $\mathfrak{R}_{i, j+1}$.

We shall now show that if \mathfrak{M}_{ij} is a manifold of discontinuity of precisely one of the functions Y^* , Z^* , then W_t and W_x are continuous at points of \mathfrak{M}_{ij} ; that is, $W_t^+ = W_t^-$ and $W_x^+ = W_x^-$ at such points. For definiteness, suppose \mathfrak{M}_{ij} is a manifold of discontinuity of Z^* but not of Y^* . The argument will be the same as that used to prove a corresponding result in [3]. At this point we remind the reader of the saddle point restriction (2.8), with particular emphasis on the nontangency requirement. Thus our assertion that $W_t^+ = W_t^-$ and $W_x^+ = W_x^-$ at points of \mathfrak{M}_{ij} is not contradicted by examples in which W_x and W_t are discontinuous at points of \mathfrak{M}_{ij} and in which the vector field is tangent at one side to \mathfrak{M}_{ij} .

Let (τ, ξ) be a point of \mathfrak{M}_{ij} . Let $t = t(s)$, and $x = x(s)$ define a $C^{(1)}$

curve C lying in \mathfrak{M}_{ij} , where $-1 \leq s \leq 1$ and $x(0) = \xi$, $t(0) = \tau$. We consider $w(s) = W(t(s), x(s))$. Since \mathfrak{M}_{ij} is of class $C^{(1)}$ and W_t^+ , W_x^+ and W^+ are continuous on $\mathfrak{R}_{i,j+1}$, it follows that there exists a $C^{(1)}$ function V defined in a neighborhood N of (τ, ξ) such that for (t, x) in $\tilde{R}_{i,j+1} \cap N$ we have $V(t, x) = W(t, x)$. Hence we have

$$w(s) = W(t(s), x(s)) = V(t(s), x(s)).$$

Therefore $w'(0)$ exists and

$$\begin{aligned} w'(0) &= V_t(\tau, \xi)t'(0) + V_x(\tau, \xi)x'(0) \\ &= W_t^+(\tau, \xi)t'(0) + W_x^+(\tau, \xi)x'(0). \end{aligned}$$

By similar arguments we get

$$w'(0) = W_t^-(\tau, \xi)t'(0) + W_x^-(\tau, \xi)x'(0).$$

Hence

$$(W_t^+ - W_t^-)t'(0) + (W_x^+ - W_x^-)x'(0) = 0.$$

Since C is an arbitrary curve and \mathfrak{M}_{ij} is an n -dimensional manifold, it follows that either

$$(4.9) \quad W_t^+ = W_t^- \quad \text{and} \quad W_x^+ = W_x^-,$$

or $(W_t^+ - W_t^-, W_x^+ - W_x^-)$ is a nonzero vector orthogonal to \mathfrak{M}_{ij} at (τ, ξ) . We shall now show that the latter is not possible, and so (4.9) must hold, which is the desired result.

Let Z_{ij}^* and $Z_{i,j+1}^*$ be extensions of Z^* as in §3. We note that if we define a strategy Z^+ by the condition that $Z^+ = Z_{i,j+1}^*$ in a neighborhood of (τ, ξ) and Z^* elsewhere, then Z^+ is in \mathfrak{Z}_1 . Moreover,

$$z^+ \equiv Z^+(\tau, \xi) = Z_{i,j+1}^*(\tau, \xi) \equiv z^{*+}.$$

Also, a strategy Z^- defined by redefining Z^* to be Z_{ij}^* in a neighborhood of (τ, ξ) is in \mathfrak{Z}_1 and

$$z^- \equiv Z^-(\tau, \xi) = Z_{ij}^*(\tau, \xi) \equiv z^{*-}.$$

We also point out that since \mathfrak{M}_{ij} is not a manifold of discontinuity of Y^* , we have

$$y^* \equiv Y^*(\tau, \xi) = Y_{ij}^*(\tau, \xi) = Y_{i,j+1}^*(\tau, \xi).$$

From the observations of the preceding paragraph and from the remarks concerning the validity of (4.7) and (4.8) at manifolds of discontinuity, we get

$$\begin{aligned} -W_t^+ &= f^+ + W_x^+ G^+ \leq f(\tau, \xi, y^*, z^-) + W_x(\tau, \xi)G(\tau, \xi, y^*, z^-) \\ &= f^- + W_x^+ G^-, \end{aligned}$$

where the f^\pm and G^\pm are evaluated at (τ, ξ, y^*, z^\pm) . Similarly,

$$\begin{aligned} -W_t^- &= f^- + W_x^- G^- \leq f(\tau, \xi, y^*, z^+) + W_x(\tau, \xi) G(\tau, \xi, y^*, z^+) \\ &= f^+ + W_x^- G^+. \end{aligned}$$

Therefore,

$$W_t^+ - W_t^- = f^- + W_x^- G^- - f^+ - W_x^+ G^+ \leq -(W_x^+ - W_x^-) G^+$$

and

$$W_t^+ - W_t^- \geq -(W_x^+ - W_x^-) G^-.$$

Hence we get

$$(4.10) \quad (W_t^+ - W_t^-) + (W_x^+ - W_x^-) G^+ \leq 0,$$

$$(4.11) \quad (W_t^+ - W_t^-) + (W_x^+ - W_x^-) G^- \geq 0.$$

If we assume that $(W_t^+ - W_t^-, W_x^+ - W_x^-)$ is different from zero and is orthogonal to \mathfrak{M}_{ij} , then the inequality must hold in both (4.10) and (4.11). For, if equality held in (4.10), say, we would have that $(1, G^-) = (1, \phi^{*'}(t_{ij} - 0, \tau, \xi))$ is tangent to \mathfrak{M}_{ij} , contrary to assumption. On the other hand, if the inequality holds in both (4.10) and (4.11) we would have the two one-sided tangent vectors pointing into opposite half-spaces of the tangent planes to \mathfrak{M}_{ij} at (τ, ξ) , which is impossible. Hence (4.9) must hold.

We summarize the principal results of this section in the following theorem.

THEOREM 1. *The value function W is continuous on \mathfrak{R} . On each \mathfrak{R}_{ij} the functions W_t and W_x exist, are continuous, and have continuous extensions to \mathfrak{R}_{ij} . If \mathfrak{M}_{ij} is a manifold of discontinuity of only one of the functions Y^* or Z^* , then W_t and W_x are continuous at points of \mathfrak{M}_{ij} . The function W satisfies (4.7) and (4.8) at all points of $\mathfrak{R} \cup \mathfrak{T}$, provided we interpret W_t , W_x , Y^* and Z^* as the appropriate limits at points of \mathfrak{T} , at points of manifolds $\mathfrak{R}_{i_1, \dots, i_k}$, and manifolds of discontinuity of both Y^* and Z^* . At a manifold \mathfrak{M}_{ij} of discontinuity of only one of the functions Y^* or Z^* , say Z^* for definiteness, (4.7) and (4.8) hold for $Z^*(t, x) = Z_{ij}^*(t, x)$ and $Z^*(t, x) = Z_{i,j+1}^*(t, x)$.*

The form of (4.5) can be recast as follows. Define

$$(4.12) \quad H(t, x, y, z, \lambda) = f(t, x, y, z) + \lambda G(t, x, y, z).$$

Then (4.4) can be written in the form

$$(4.13) \quad H(t, x, Y^*(t, x), Z^*(t, x), W_x(t, x)) + W_t(t, x) = 0.$$

Thus we have the following corollary to Theorem 1.

COROLLARY. *W satisfies the Hamilton-Jacobi equation (4.13) on \mathfrak{R} , where*

we give the usual interpretations to Y^* , Z^* , W_x , and W_t at manifolds of discontinuity.

5. The adjoint variables or Lagrange multipliers. In this section we shall introduce the adjoint variables or Lagrange multipliers, and study their relationship to W_t and W_x .

Consider the function W on \mathfrak{R}_{ij} . It follows from (4.1) and Lemma 3 that W_ξ is given by the following formula:

$$(5.1) \quad \begin{aligned} W_\xi(\tau, \xi) = & g_\sigma \frac{\partial \sigma}{\partial \xi} + f(P_{ij_i}) \frac{\partial t_{ij_i}}{\partial \xi} \\ & + \sum_{k=j}^{j_i-1} [f(P_{ik} - 0) - f(P_{ik} + 0)] \frac{\partial t_{ik}}{\partial \xi} \\ & + \left(\int_\tau^{t_{ij}} + \sum_{k=j}^{j_i-1} \int_{t_{ik}}^{t_{i,k+1}} \right) (\tilde{f}_x + \tilde{f}_y \tilde{Y}_x^* + \tilde{f}_z \tilde{Z}_x^*) \phi_\xi^* dt, \end{aligned}$$

where \tilde{f}_x , \tilde{f}_y , and \tilde{f}_z are evaluated as in (4.2), $P_{ik} + 0$ is given by (3.9) with j replaced by k , $P_{ik} - 0$ is given by (3.10) with j replaced by k , $P_{ij_i} = P_{ij_i} - 0$, and

$$\tilde{Y}_x^* = Y_x^*(t, \phi^*(t, \tau, \xi)),$$

$$\tilde{Z}_x^* = Z_x^*(t, \phi^*(t, \tau, \xi)),$$

$$g_\sigma = \frac{\partial g}{\partial t}(t_{ij_i}, x_{ij_i}) \frac{\partial t_{ij_i}}{\partial \sigma} + \frac{\partial g}{\partial x}(t_{ij_i}, x_{ij_i}) \frac{\partial x_{ij_i}}{\partial \sigma}.$$

We shall simplify (5.1) presently.

Consider the following system of n linear equations in the n unknown components of λ_{j_i} :

$$(5.2) \quad g_\sigma + f(P_{ij_i}) \frac{\partial T_{ij_i}}{\partial \sigma} + \lambda_{j_i} \left(G(P_{ij_i}) \frac{\partial T_{ij_i}}{\partial \sigma} - \frac{\partial \chi_{ij_i}}{\partial \sigma} \right) = 0,$$

where g_σ , $\partial T_{ij_i}/\partial \sigma$, and $\partial \chi_{ij_i}/\partial \sigma$ are evaluated at the parameter value of σ corresponding to $t_{ij_i}(\tau, \xi)$, $x_{ij_i}(\tau, \xi)$, i.e., the endpoint of the optimal path from (τ, ξ) . If we substitute $\phi^{*'}(t_{ij_i} - 0, \tau, \xi)$ for $G(P_{ij_i})$ in (5.2) and use the properties of f and Lemma 3, we see that (5.2) defines λ_{j_i} uniquely as a continuous function of (τ, ξ) on \mathfrak{R}_{ij} .

In terms of the function H introduced in (4.12) we can write (5.2) as follows:

$$(5.3) \quad g_\sigma + H \frac{\partial T_{ij_i}}{\partial \sigma} - \lambda_{j_i} \frac{\partial \chi_{ij_i}}{\partial \sigma} = 0,$$

where H is evaluated at $(t_{ij_i}, x_{ij_i}, y^*(t_{ij_i}), z^*(t_{ij_i}), \lambda_{j_i})$. If we multiply

(5.3) on the right by $\partial\sigma/\partial\xi$ and use (3.14) of Lemma 3, we obtain

$$(5.4) \quad g_\sigma \frac{\partial\sigma}{\partial\xi} + H \frac{\partial t_{ij_i}}{\partial\xi} - \lambda_{j_i} \frac{\partial x_{ij_i}}{\partial\xi} = 0.$$

On the interval $t_{i,j-1}(\tau, \xi) \leq t \leq t_{ij_i}(\tau, \xi)$ consider the differential equation

$$(5.5) \quad \frac{d\lambda}{dt} = -(\bar{H}_x + \bar{H}_y \bar{Y}_x^* + \bar{H}_z \bar{Z}_x^*),$$

subject to the initial conditions

$$\lambda(t_{ij_i}) = \lambda_{j_i},$$

where the bar indicates that the arguments of the functions f, G are evaluated as in (4.2), where \bar{Y}_x^* and \bar{Z}_x^* are as in (5.1), and where λ_{j_i} is given by (5.2) or (5.4). Since (5.5) is a linear system in λ it follows that (5.5) has a unique solution satisfying the initial conditions on the interval $[t_{i,j-1}, t_{ij_i}]$. Furthermore, since λ_{j_i} and t_{ij_i} are continuous functions of (τ, ξ) for (τ, ξ) in \mathfrak{R}_{ij} , it follows from standard theorems that we may write the solution as

$$\lambda = \lambda(t, t_{ij_i}, \lambda_{j_i}) = \lambda(t, \tau, \xi),$$

where $\lambda(t, \tau, \xi)$ is a continuous function of (t, τ, ξ) for (τ, ξ) in \mathfrak{R}_{ij} and $t_{i,j-1} \leq t \leq t_{ij_i}$.

Set $\lambda_{j_{i-1}}^+ = \lambda(t_{i,j_{i-1}} + 0, \tau, \xi)$ and consider the following linear system in the unknown $\lambda_{j_{i-1}}^-$:

$$(5.6) \quad \begin{aligned} & -\lambda_{j_{i-1}}^- \left[G(P_{i,j_{i-1}} - 0) \frac{\partial T_{i,j_{i-1}}}{\partial\sigma} - \frac{\partial \chi_{i,j_{i-1}}}{\partial\sigma} \right] \\ & = [f(P_{i,j_{i-1}} - 0) - f(P_{i,j_{i-1}} + 0)] \frac{\partial T_{i,j_{i-1}}}{\partial\sigma} \\ & \quad - \lambda_{j_{i-1}}^+ \left[G(P_{i,j_{i-1}} + 0) \frac{\partial T_{i,j_{i-1}}}{\partial\sigma} - \frac{\partial \chi_{i,j_{i-1}}}{\partial\sigma} \right], \end{aligned}$$

where $\partial T_{i,j_{i-1}}/\partial\sigma$ and $\partial \chi_{i,j_{i-1}}/\partial\sigma$ are evaluated at the point $(t_{i,j_{i-1}}, x_{i,j_{i-1}})$ at which $\phi^*(t, \tau, \xi)$ intersects $\mathfrak{M}_{i,j_{i-1}}$. From the relation

$$\phi^{*'}(t_{i,j_{i-1}} + 0, \tau, \xi) = G(P_{i,j_{i-1}} + 0),$$

from Lemma 3, and the continuity in (τ, ξ) of $\lambda(t_{i,j_{i-1}} - 0, \tau, \xi)$, it follows that (5.6) defines $\lambda_{j_{i-1}}^-$ uniquely as a continuous function of (τ, ξ) for (τ, ξ) in \mathfrak{R}_{ij} . It therefore follows that (5.5) subject to the initial conditions

$$\lambda(t_{i,j_{i-1}}) = \lambda_{j_{i-1}}^-$$

has a unique solution on the interval $t_{i,j_i-2} \leq t \leq t_{i,j_i-1}$. Moreover, for t in this interval and (τ, ξ) in \mathcal{R}_{ij} , the solution $\lambda(t, \tau, \xi)$ is a continuous function of (t, τ, ξ) .

From the definition of H it is clear that an alternate way of writing (5.6) is the following:

$$(5.7) \quad \begin{aligned} H(\Pi_{i,j_i-1}^-) \frac{\partial T_{i,j_i-1}}{\partial \sigma} - \lambda_{j_i-1}^- \frac{\partial \chi_{i,j_i-1}}{\partial \sigma} \\ = H(\Pi_{i,j_i-1}^+) \frac{\partial T_{i,j_i-1}}{\partial \sigma} - \lambda_{j_i-1}^+ \frac{\partial \chi_{i,j_i-1}}{\partial \sigma}, \end{aligned}$$

where

$$\Pi_{i,k}^\pm = (t_{ik}, x_{ik}, Y^{*\pm}(t_{ik}, x_{ik}), Z^{*\pm}(t_{ik}, x_{ik}), \lambda_k^\pm),$$

for $k = j, \dots, j_i - 1$. Here, $Y^{*+} = Y_{i,k+1}^*$, $Y^{*-} = Y_{i,k}^*$, and similarly for $Z^{*\pm}$. For $k = j_i$, we define Π_{i,j_i} as above, except that we delete the \pm signs. If we multiply (5.7) through on the right by $\partial \sigma / \partial \xi$ and use (3.14) of Lemma 3, we get

$$H(\Pi_{i,j_i-1}^-) \frac{\partial t_{i,j_i-1}}{\partial \xi} - \lambda_{j_i-1}^- \frac{\partial x_{i,j_i-1}}{\partial \xi} = H(\Pi_{i,j_i-1}^+) \frac{\partial t_{i,j_i-1}}{\partial \xi} - \lambda_{j_i-1}^+ \frac{\partial x_{i,j_i-1}}{\partial \xi}.$$

We can proceed backwards in time in this fashion across each of the manifolds $\mathcal{M}_{i,j_i-1}, \dots, \mathcal{M}_{ij}$, and obtain the following result.

LEMMA 4. *There exists a function $\lambda(t, \tau, \xi)$ defined for (τ, ξ) in \mathcal{R}_{ij} and all $t_{i,k-1} < t < t_{ik}$, $k = j, j+1, \dots, j_i$, such that the following hold:*

- (i) *For (τ, ξ) in \mathcal{R}_{ij} and $t_{i,k-1} < t < t_{ik}$, λ is a continuous function of (t, τ, ξ) .*
- (ii) *As $t \rightarrow t_{ik} + 0$ and $t \rightarrow t_{ik} - 0$, $k = j, \dots, j_i - 1$, and as $t \rightarrow t_{ij_i} - 0$, the function λ tends to unique limits*

$$\begin{aligned} \lambda_k^- &= \lambda(t_{ik} - 0, \tau, \xi), & \lambda_k^+ &= \lambda(t_{ik} + 0, \tau, \xi) \\ \lambda_{j_i} &= \lambda(t_{ij_i}, \tau, \xi). \end{aligned}$$

Moreover λ_k^- and λ_k^+ satisfy

$$(5.8) \quad H(\Pi_{ik}^-) \frac{\partial t_{ik}}{\partial \xi} - \lambda_k^- \frac{\partial x_{ik}}{\partial \xi} = H(\Pi_{ik}^+) \frac{\partial t_{ik}}{\partial \xi} - \lambda_k^+ \frac{\partial x_{ik}}{\partial \xi},$$

for $k = j, \dots, j_i - 1$, and λ_{j_i} satisfies (5.4).

- (iii) *On each interval $[t_{i,k-1}, t_{ik}]$, $k = j, \dots, j_i$, $\lambda'(t, \tau, \xi)$ exists and satisfies (5.5), where at the endpoints λ and its derivatives are taken as limits from the interior.*

We note one more simple, but important, relation. Let $\hat{\xi} = \phi^*(\hat{\tau}, \tau, \xi)$, where $\hat{\tau} > \tau$. Then for $t \geq \hat{\tau}$, $\phi^*(t, \hat{\tau}, \hat{\xi}) = \phi^*(t, \tau, \xi)$. It therefore follows

that for $t \geq \hat{\tau}$,

$$(5.9) \quad \lambda(t, \tau, \xi) = \lambda(t, \hat{\tau}, \hat{\xi}).$$

Let us now return to (5.1). Let λ be the function in Lemma 4. Then, since λ satisfies (5.5) we may rewrite the terms involving the integrals in (5.1) as follows

$$(5.10) \quad \left(\int_{\tau}^{t_{ij}} + \sum_{k=j}^{j_i-1} \int_{t_{ik}}^{t_{i,k+1}} \right) \left(-\frac{d\lambda}{dt} - \lambda \bar{G}_y \bar{Y}_x^* - \lambda \bar{G}_z \bar{Z}_x^* \right) \phi_{\xi}^* dt.$$

From the relation

$$\phi^{*'}(t, \tau, \xi) = G(t, \phi^*(t, \tau, \xi), Y^*(t, \phi^*(t, \tau, \xi)), Z^*(t, \phi^*(t, \tau, \xi)))$$

and from Lemma 3 it follows that for $t_{i,k-1} < t < t_{ik}$, $k = j, j+1, \dots, j_i$,

$$\phi_{t\xi}^*(t, \tau, \xi) \equiv \phi_{\xi}^{*'}(t, \tau, \xi) = (\bar{G}_x + \bar{G}_y \bar{Y}_x^* + \bar{G}_z \bar{Z}_x^*) \phi_{\xi}^*.$$

It further follows from Lemma 3 that for $t_{i,k-1} < t < t_{ik}$, $k = j, \dots, j_i$,

$$\phi_{t\xi}^*(t, \tau, \xi) = \phi_{\xi t}^*(t, \tau, \xi).$$

Hence (5.10) is equal to

$$(5.11) \quad \begin{aligned} & - \left(\int_{\tau}^{t_{ij}} + \sum_{k=j}^{j_i-1} \int_{t_{ik}}^{t_{i,k+1}} \right) d(\lambda \phi_{\xi}^*) \\ & = -\lambda_{j_i} \phi_{\xi}^*(t_{ij_i}, \tau, \xi) + \lambda(\tau, \tau, \xi) \phi_{\xi}^*(\tau, \tau, \xi) \\ & \quad + \sum_{k=j}^{j_i-1} [\lambda_k^+ \phi_{\xi}^*(t_{ik} + 0, \tau, \xi) - \lambda_k^- \phi_{\xi}^*(t_{ik} - 0, \tau, \xi)]. \end{aligned}$$

From Lemma 3 we have that the one-sided limits of ϕ_{ξ}^* at $t = t_{ik}$, $k = j, j+1, \dots, j_i$, are given by (3.10) and (3.12) with j replaced by k . Also, from standard theorems in the theory of ordinary differential equations we have $\phi_{\xi}^*(\tau, \tau, \xi) = I$. If we make these substitutions into the right-hand side of (5.11) and then use (5.8) and (5.4), we get

$$(5.12) \quad W_{\xi}(\tau, \xi) = \lambda(\tau, \tau, \xi).$$

Let $t > \tau$, $t \neq t_{ik}$, and let $x = \phi^*(t, \tau, \xi)$. Then by (5.12) and (5.9),

$$(5.13) \quad W_x(t, x) = \lambda(t, t, x) = \lambda(t, \tau, \xi).$$

Since W_x is continuous at manifolds of discontinuity of only one of the functions Y^* and Z^* , it follows that for fixed (τ, ξ) , λ is continuous at values t_{ik} corresponding to manifolds of discontinuity of only one of the functions Y^* or Z^* .

Let (t, x) be a point on the optimal path from (τ, ξ) ; thus $x = \phi^*(t, \tau, \xi)$.

It then follows from (4.8) and (5.13) that

$$\begin{aligned}
 \max_y \min_z [f + \lambda G] &= \max_y \min_z H(t, x, y, z, \lambda) \\
 (5.15) \qquad &= \min_z \max_y H(t, x, y, z, \lambda) \\
 &= H(t, x, y^*, z^*, \lambda) = -W_t(t, x),
 \end{aligned}$$

where $x = \phi(t, \tau, \xi)$, $\lambda = \lambda(t, \tau, \xi)$ and

$$\begin{aligned}
 y &\in E[y \mid y = Y(t, x), Y \in \mathfrak{Y}_1], \\
 z &\in E[z \mid z = Z(t, x), Z \in \mathfrak{Z}_1].
 \end{aligned}$$

At points $t = t_{ik}$, (5.15) holds for the one-sided limits.

We summarize the principal results of this section in the following theorem.

THEOREM 2. *Let $\phi^*(t, \tau, \xi)$ be the optimal path from a point (τ, ξ) in \mathfrak{R}_{ij} . Then there exists a function $\lambda(t, \tau, \xi)$ defined for $\tau \leq t \leq t_{ij_i}(\tau, \xi)$ and $t \neq t_{ik}$, $k = j, j+1, \dots, j_i-1$, such that the following hold:*

(i) *λ is continuous on its domain of definition, and at the points t_{ik} possesses a right-hand limit λ_k^+ and a left-hand limit λ_k^- .*

(ii) *The functions λ and ϕ^* satisfy the following system of differential equations:*

$$\begin{aligned}
 \frac{dx}{dt} &= H_\lambda(t, x, Y^*(t, x), Z^*(t, x), \lambda), \\
 (5.17) \qquad \frac{d\lambda}{dt} &= -\frac{\partial H}{\partial x}(t, x, Y^*(t, x), Z^*(t, x), \lambda),
 \end{aligned}$$

where

$$\frac{\partial H}{\partial x} = H_x + H_y Y_x^* + H_z Z_x^*,$$

and at $t = t_{ik}$, $k = j, j+1, \dots, j_i-1$, the equations hold for the one-sided limits.

(iii) *If \mathfrak{M}_{ik} , $j \leq k \leq j_i-1$, is a manifold of discontinuity of only one of the functions Y^*, Z^* , then λ is continuous at $t = t_{ik}$, i.e., $\lambda_k^+ = \lambda_k^-$. Otherwise, (5.8) holds.*

(iv) *At $t = t_{ij_i}$ the transversality relation (5.4) holds.*

(v) *If $x = \phi^*(t, \tau, \xi)$, $t \geq \tau$, then*

$$W_x(t, x) = \lambda(t, \tau, \xi).$$

(vi) *For all $\tau \leq t \leq t_{ij_i}$ and $t \neq t_{ik}$, $k = j, \dots, j_i-1$,*

$$\begin{aligned}
 \max_y \min_z H(t, \phi^*(t), y, z, \lambda(t)) &= \min_z \max_y H(t, \phi^*(t), y, z, \lambda(t)) \\
 &= H(t, \phi^*(t), y^*(t), z^*(t), \lambda(t)),
 \end{aligned}$$

where $y^*(t) = Y^*(t, \phi^*(t))$, $Z^*(t) = Z^*(t, \phi^*(t))$, and y and z are as in (5.16). If $t = t_{ik}$, $k = j, \dots, j_i - 1$, the above holds for the one-sided limits.

Remark. By the corollary to Theorem 1, W satisfies the Hamilton-Jacobi equation (4.12). The equations (5.17) are readily seen to be the characteristic equations of (4.12).

Let us now assume that $\Omega(t, x)$ and $\bar{\Omega}(t, x)$ are defined by systems of inequalities. More precisely, let $R(t, x, z)$ be a $C^{(1)}$ function with domain in (t, x, z) -space and range in E^p . A vector z is in $\Omega(t, x)$ if and only if $R(t, x, z) \geq 0$. Similarly, we suppose that $\bar{\Omega}(t, x)$ is defined by $K(t, x, y) \geq 0$, where K is a $C^{(1)}$ function with range in \bar{p} -dimensional Euclidean space. We suppose that R and K satisfy the following constraint condition.

Constraint condition. If $p > s$ (recall, s is the dimension of z), then at each point (t, x, z) , at most s components of R can vanish. The matrix $\partial R^i / \partial z^j$ formed from those components R^i of R that vanish at (t, x, z) has maximum rank at (t, x, z) . A similar statement holds for K .

For constraints of the form just described we have the following corollary to Theorem 2.

COROLLARY 1. *Let $\Omega(t, x)$ be given by a system of inequalities $R(t, x, z) \geq 0$ and let $\bar{\Omega}(t, x)$ be given by a system of inequalities $K(t, x, y) \geq 0$, where R and K satisfy the constraint condition. Then there exist functions $\mu(t, \tau, \xi)$ and $\nu(t, \tau, \xi)$ defined for $\tau \leq t \leq t_{ij}$ and $t \neq t_{ik}$ such that the following holds. At all points $(t, \phi^*(t), y^*(t), z^*(t), \lambda(t))$, where $y^*(t) = Y^*(t, \phi^*(t), z^*(t)) = Z^*(t, \phi^*(t))$,*

$$(5.18) \quad H_y + \nu K_y = 0, \quad H_z + \mu R_z = 0,$$

$$(5.19) \quad \nu^j K^j = 0, \quad j = 1, \dots, \bar{p}, \quad \mu^i R^i = 0, \quad i = 1, \dots, p,$$

$$(5.20) \quad \nu \geq 0, \quad \mu \leq 0.$$

Proof. Conclusion (vi) of Theorem 2 implies that for $\tau \leq t \leq t_{ij}$,

$$(5.21) \quad \begin{aligned} H(t, \phi^*(t), y, z^*(t), \lambda(t)) &\leq H(t, \phi^*(t), y^*(t), z^*(t), \lambda(t)) \\ &\leq H(t, \phi^*(t), y^*(t), z, \lambda(t)) \end{aligned}$$

for all y and z as in (5.16), with the appropriate interpretation at $t = t_{ik}$.

Let us fix t_0 , $t_0 \neq t_{ik}$, $k = j, \dots, j_i$, and let z_0 satisfy $R(t_0, \phi^*(t_0), z_0) \geq 0$. Let \hat{R} be the vector consisting of those components R^i of R such that $R^i(t_0, \phi^*(t_0), z_0) = 0$. Thus,

$$(5.22) \quad \hat{R}(t, x, z) = 0$$

has a solution $(t_0, \phi^*(t_0), z_0)$. From the constraint condition and the implicit function theorem it follows that there are a neighborhood N of $(t_0, \phi^*(t_0))$ and a $C^{(1)}$ function $Z_0(t, x)$ defined on this neighborhood such that

$R(t, x, Z_0(t, x)) = 0$ on this neighborhood. Moreover, since for these components of R not in \hat{R} we have $R^i(t_0, \phi^*(t_0), z_0) > 0$, it follows by continuity that we may restrict the neighborhood N so that $R(t, x, Z_0(t, x)) \geq 0$ on this neighborhood. It is not hard to see that N may be further restricted so that the strategy Z defined by

$$Z(t, x) = \begin{cases} Z_0(t, x), & (t, x) \in N, \\ Z^*(t, x), & (t, x) \notin N, \end{cases}$$

is in \mathfrak{Z}_1 .

To summarize, we have shown that for any z such that $R(t_0, \phi^*(t_0), z) \geq 0$, there is a strategy Z in \mathfrak{Z}_1 such that $z = Z(t_0, \phi^*(t_0))$. Since t_0 is arbitrary, $t_0 \neq t_{ik}$, this holds for all $t \neq t_{ik}$. Therefore, it follows that for all $\tau \leq t \leq t_{ij}$ and $t \neq t_{ik}$, the second inequality in (5.21) holds for all z such that $R(t, \phi^*(t), z) \geq 0$. In other words, $z^*(t)$ minimizes $H(t, \phi^*(t), y^*(t), z, \lambda(t))$ subject to $R(t, \phi^*(t), z) \geq 0$.

It is not hard to see that if our constraint condition holds, then the Kuhn-Tucker constraint qualification holds [7]. Therefore, at each $t \neq t_{ik}$ we can apply the Kuhn-Tucker necessary conditions [7] and obtain the existence of $\mu(t, \tau, \xi)$ satisfying (5.18)–(5.20). Arguments similar to the ones just presented give the existence of a $\nu(t, \tau, \xi)$ satisfying (5.18)–(5.20).

If we make one more mild assumption, we can sharpen Corollary 1 and eliminate the functions Z_x^* and Y_x^* from (5.17).

COROLLARY 2. *Let (τ, ξ) be a point of \mathfrak{R}_{ij} . Suppose that on each interval $t_{i,k-1} \leq t \leq t_{ik}$, $k = j, \dots, j_i$, the components R^i of R such that $R^i(t, \phi^*(t, \tau, \xi), z^*(t)) = 0$ do not change, and suppose that a similar statement holds for K . Then μ and ν are continuous functions of (t, τ, ξ) for (τ, ξ) in \mathfrak{R}_{ij} and $t_{i,k-1} < t < t_{ik}$, and have one-sided limits at the endpoints $t_{i,k-1}$ and t_{ik} . At the endpoints t_{ik} , (5.18)–(5.20) hold for the one-sided limits.*

The second equation in (5.17) can be replaced by the equation

$$(5.23) \quad \frac{d\lambda}{dt} = -H_x - \mu R_x - \nu K_y.$$

Proof. Let us fix our attention on an interval $t_{i,k-1} < t < t_{ik}$. From (5.19) it follows that if $R^i > 0$, then $\mu^i = 0$. Hence these components of μ are continuous for $t_{i,k-1} < t < t_{ik}$ and have one-sided limits. We may therefore write (5.18) in the form $H_z + \hat{\mu} \hat{R}_z = 0$. It now follows from the constraint condition that we can select an appropriate nonsingular submatrix of R_z , which we denote by M , and write $\mu = M^{-1} h_z$, where h_z is a column vector made up of the appropriate entries from H_z . Since the entries of M^{-1} and h_z are continuous functions of (t, τ, ξ) for $t_{i,k-1} < t < t_{ik}$ and (τ, ξ) in \mathfrak{R}_{ij} and have one-sided limits, it follows that the same is true for μ . Hence the statement that μ is continuous and has one-sided limits is proved. A similar argu-

ment holds for ν . The statement that (5.18)–(5.20) hold for the one-sided limits now follows.

To establish (5.23) we first note that from (5.18) we have

$$(5.24) \quad H_y Y_x^* = -\nu K_y Y_x^*, \quad H_z Z_x^* = -\mu R_z Z_x^*.$$

Since $\hat{R}(t, \phi^*(t), z^*(t)) = 0$ and since $\hat{R}(t, x, Z^*(t, x)) \geq 0$, it follows that each component of \hat{R} has a relative minimum at $(t, \phi^*(t))$. Hence $\hat{R}_x + \hat{R}_z Z_x^* = 0$, and so $\hat{\mu} \hat{R}_x + \hat{\mu} \hat{R}_z Z_x^* = 0$. Since those components of μ not in $\hat{\mu}$ are zero on $t_{i,k-1} \leq t \leq t_{ik}$, we have

$$\mu R_x + \mu R_z Z_x^* = 0.$$

Similarly, we obtain

$$\nu K_x + \nu K_y Y_x^* = 0.$$

Combining these last two equations with (5.24) and then substituting the result into (5.17), we get (5.23).

Remark. It is clear from the proof that the assumption that on an interval $(t_{i,k-1}, t_{ik})$ the components R^i of R such that $R^i = 0$ do not change, can be replaced by the assumption that there are a finite number of changes in the components of R .

We conclude by calling attention to one further necessary condition, namely Theorem 6 of [2]. We refer the reader to [2] for the theorem, and leave its proof in the present context for the reader. We also refer the reader to [2] for a sufficiency theorem.

REFERENCES

- [1] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.
- [2] L. D. BERKOVITZ, *A variational approach to differential games*, Advances in Game Theory, Annals of Math. Study 52, Princeton University Press, Princeton, 1964, pp. 127–174.
- [3] L. D. BERKOVITZ AND S. E. DREYFUS, *A dynamic programming approach to the nonparametric problem in the calculus of variations*, J. Math. Mech., 15 (1966), pp. 83–100.
- [4] L. M. GRAVES, *Theory of Functions of Real Variables*, 2nd ed., McGraw-Hill, New York, 1956.
- [5] R. ISAACS, *Differential games III*, Research Memorandum RM-1411, The RAND Corporation, Santa Monica, California, 1954.
- [6] ———, *Differential Games*, John Wiley, New York, 1965.
- [7] H. J. KUHN AND A. W. TUCKER, *Nonlinear programming*, Proceedings of the Second Berkeley Symposium on Mathematics and Probability, University of California Press, Berkeley and Los Angeles, 1951, pp. 481–492.

APPLICATIONS OF FUNCTIONAL ANALYSIS TO THE THEORY OF OPTIMAL PROCESSES*

F. M. KIRILLOVA†

1. Introduction. In the years of development of optimal control theory, powerful general methods were created, based on the now widely known "maximum principle" and "optimality principle." The maximum principle is the most convenient method for solving problems of optimal programmed control; a detailed exposition is found in the fundamental work of L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mishchenko [1]. The methods of dynamic programming are given for a large class of problems of control, together with computational schemes for solving functional equations, by R. Bellman and his colleagues [2]. Parallel to the development in these and other directions (see [3], [4]) starting in 1956, attempts have been made to introduce methods of functional analysis into the study of optimal control problems.

At first it seemed that the methods of functional analysis applied only to a very restricted class of problems. But in spite of this, the number of studies using the ideas of functional analysis has increased. This is apparently explained by the fact that, in the solution of optimal control problems, with the help of the maximum principle, or by reduction to the Euler equations, there remains an indeterminate last step: as is known, these methods do not show how to select the initial condition for solving the adjoint system. The methods of dynamic programming and the approach that leads to the Hamilton-Jacobi equations do not have this deficiency. However, the solution of functional equations, to which both paths lead, is not an easy problem, and the advantages of the equation $u = u(x)$ over $u = u(x(t_0), t)$ are open to dispute if the control system is subject to the influence of perturbing forces or is nonstationary.

The functional approach to problems of optimal programmed control, which is described below, reduces variational problems to operations with functions of a finite number of variables. These are, as a rule, convex or concave functions, and determination of their extrema completes the solution of the optimal control problem. Game situations arise in many cases (statistical problems, pursuit problems, etc.), and the problem of determining the control functions reduces to the solution of a game whose players have finite-dimensional vectors as their strategies.

* Received by the editors June 1, 1966.

This translation into English has been prepared by Lisa Rosenblatt and supported through a grant-in-aid from the National Science Foundation.

† Scientific Research Department of Power and Automation, S. M. Kirov Ural Polytechnic Institute, Sverdlovsk 2, USSR.

The method using the ideas of functional analysis turned out to be very fruitful for a complete study of a wide class of problems in the theory of optimal processes (deterministic, statistical, adaptive), whether time-optimal, or optimal in the sense of terminal error or other criteria, as well as for two-point problems, both with variable endpoints and free endpoints. One of the typical features of the method is that it yields necessary and sufficient conditions for the existence of solutions. This fact makes it possible to study qualitative aspects of optimal processes: questions of controllability, existence and uniqueness of optimal controls, continuous dependence of solutions on initial data or parameters, mutual dependence of solutions of problems with various types of restrictions, etc.

Below we give a survey of certain works where the methods of functional analysis are used in solving problems in the theory of optimal processes. We describe methods for reducing variational (infinite-dimensional) problems to operations with functions of a finite number of variables, and we give computational algorithms; in addition we give conditions for controllability and existence of solutions for certain optimal problems.

It is not the author's intention to give a complete exposition of the problem of applications of functional analysis to the theory of optimal processes. In particular, she does not touch on the work of A. Ya. Dubovitskii and A. A. Milyutin, where the application of functional analysis to optimal control problems is treated differently. The main concern is with results obtained by the author and her colleagues in the Scientific Research Department of Power and Automation of the Ural Polytechnic Institute. In passing, results are also given from the work of other authors relating to the compatibility of the approach to the study of optimal processes.

We discuss basically optimal processes for objects described by ordinary linear differential equations, although the methods of solution may be extended to integrodifferential equations, partial differential equations, and certain nonlinear systems. We note that, from the point of view of concrete computations, generalizations to nonlinear equations are usually not as efficient as in the linear case. But since computation of optimal processes in nonlinear systems is often based on successive linearizations or piecewise-linear approximations, and leads to solution of the corresponding problems for linear systems, a complete study of linear systems from the functional analysis viewpoint is of interest, also, as a first step in analogous problems for nonlinear systems.

After describing in §2 the control systems on which the study is based, we discuss the following questions in §§3–10:

1. The aspects of functional analysis applicable to the theory of optimal processes (§3).
2. Problems of controllability of linear systems (§4).

3. Examples of the reduction of control problems to functional problems. Basic relations in optimal systems (§5).
4. Certain statistical problems of optimal control (§6).
5. Continuous dependence of solutions of optimal control problems on initial data and parameters (§7).
6. Problems of numerical solution (§7).
7. Application of functional analysis to certain problems of pursuit (§9).
8. Possible generalizations (§10).

2. Basic control systems. Consider the vector equations:

$$(1) \quad \frac{dx}{dt} = A(t)x + C(t)u + f(t)$$

and

$$(2) \quad \frac{dx}{dt} = A(t)x(t) + B(t)x(t-h) + C(t)u(t), \quad h > 0,$$

where $x = (x_1, \dots, x_n)$, $x \in X$, X is an n -dimensional space, $A(t)$, $B(t)$, $C(t)$ are matrices of dimension $n \times n$, $n \times n$, $n \times r$, respectively, $f(t) = (f_1(t), \dots, f_n(t))$, $f_i(t)$ are external perturbations, u is an r -dimensional control function, h is a constant delay.

The solution of (1) at time $t = T$ for a given function $u = u(\tau)$ can be written

$$x(u, T, t_0) = x(T) = F(T, t_0)x(t_0) + \int_{t_0}^T F(T, \tau) [C(\tau)u(\tau) + f(\tau)] d\tau,$$

where t_0 is the initial moment and the matrix $F(t, \tau)$ satisfies the conditions

$$\frac{\partial F(t, \tau)}{\partial t} = A(t)F(t, \tau), \quad F(t_0, t_0) = E,$$

E is the identity matrix.

Letting

$$\int_{t_0}^T F(T, \tau)C(\tau)u(\tau) d\tau = Su,$$

$$F(T, t_0)x(t_0) = c_1,$$

$$\int_{t_0}^T F(T, \tau)f(\tau) d\tau = c_2,$$

we arrive at an operator form of the solution to (1), which we shall also use later:

$$(3) \quad x(T) = x = Su + c, \quad c = c_1 + c_2.$$

We consider next (2). Suppose we are given the function $\phi(\tau)$, and $x(\tau) \equiv \phi(\tau)$, $t_0 - h \leq \tau < t_0$. We can show that

$$(4) \quad x = S_1 u + c^1,$$

where S_1 is a linear operator, c^1 is a constant vector. If (2) is a stationary system, then

$$S_1 u = \int_{t_0}^T F(T, \tau) C u(\tau) d\tau,$$

$$\frac{\partial F(t, \tau)}{\partial t} = A F(t, \tau) + B F(t - h, \tau), \quad F(t, \tau) \equiv 0 \quad \text{for } \tau > t,$$

$$\lim_{\tau \rightarrow t+0} F(t, \tau) = 0, \quad \lim_{\tau \rightarrow t-0} F(t, \tau) = E,$$

$$c^1 = \int_{t_0-h}^{t_0} F(T, \tau + h) B \phi(\tau) d\tau + F(T, t_0) x(t_0).$$

Let W be a finite-dimensional normed linear space. We give a few well-known definitions.

DEFINITION 1. The function $\gamma = \gamma(w)$ is said to be *convex* in the convex region Δ , $\Delta \subset W$, if for all $w_1, w_2 \in \Delta$ and $\alpha \in [0, 1]$ we have the inequality

$$\gamma(\alpha w_1 + (1 - \alpha)w_2) \leq \alpha \gamma(w_1) + (1 - \alpha)\gamma(w_2).$$

DEFINITION 2. The function $\beta = \beta(w)$, $w \in \Delta$, is *quasi-convex* in w if for each δ the set $\Phi(w) = \{w: \beta(w) \leq \delta\}$ is convex.

DEFINITION 3. The hyperplane $H = \{w \in W: f(w) = 0\}$ is said to *support* the set M at the point $w_0 \in M$ if M lies on one side of H and $w_0 \in H$.

DEFINITION 4. Let X be a normed linear space. Let X^* denote the space adjoint to X . If $X^{**} = X$, then we say that the space X is *reflexive*.

3. The aspects of functional analysis applicable to the theory of optimal processes. For the exposition of control problems we state some theorems from functional analysis [5]–[8] which are basic to the solution of optimal control problems.

3.1 Theorem on the separability of closed convex sets.

THEOREM 1. Let M_1 and M_2 be disjoint closed convex subsets of the reflexive Banach space X , and let one of them be bounded. Then the sets M_1 and M_2 can be separated by a hyperplane.

Thus the theorem asserts the existence of a linear functional f , $f \in X^*$, and a number α , such that

$$f(h) \leq \alpha \quad \text{for } h \in M_1 \quad \text{and} \quad f(h) > \alpha \quad \text{for } h \in M_2.$$

This theorem was used in 1956 by R. Bellman, I. Glicksberg and O. Gross in a two-point problem of time-optimal control. The same approach was used in 1963 by H. A. Antosiewicz [10] on a problem with variable right endpoint. R. Gabasov and the author showed the possibility of using a theorem on separability of closed convex sets for other control problems [11], [12], and they discovered a clear connection between control problems and the theory of linear inequalities [13]. We will indicate still more problems to which this theorem can be applied.

Let $p_i(z)$ be quasi-convex functions of z .

PROBLEM (a). Minimization of a quasi-convex function of the final state of system (1). Given the duration of the process, $\tau = T - t_0$, determine a function $u^0(t)$ such that

$$p_1(x)|_{t=t_0} \leq 0, \quad p_2(x(u^0, T, t_0)) = \min_{u \in U} p_2(x(u, T, t_0)) = \delta^0,$$

where $U = \{u: p_3(u) \leq 0\}$.

PROBLEM (b). Transfer of an object from a given set $p_1(x)|_{t=t_0} \leq 0$ in minimum time $\tau = T^0 - t_0$ to the set $p_2(x)|_{t=T^0} \leq 0$, with $u \in U$.

PROBLEM (c). Time-optimal control for systems with delayed argument. Suppose we are given (2) and we know that $x(\tau) \equiv \phi(\tau)$ for $t_0 - h \leq \tau < t_0$, $x(t_0) = x_0$, where $\phi(\tau)$ is given a piecewise continuous function, and x_0 is a known vector. We wish to determine a control $u(t)$, $\|u\| \leq 1$, where the condition

$$x(t) \equiv 0, \quad t \geq T,$$

is guaranteed in the minimal possible time T .

In these problems, the theorem on separability of closed convex sets is used to derive sufficient conditions for existence of solutions. Thus, for example, Problem (a) is handled as follows. Let δ be a positive number. Let $\Gamma(w) = \{w: p_2(w) \leq \delta\}$, and define the set of admissibility

$$\Delta(u) = \{x: x = Su + c, u \in U, p_1(x)|_{t=t_0} \leq 0\}.$$

The sets $\Gamma(w)$, $\Delta(w)$ are convex. If $\Gamma(w)$, $\Delta(w)$ are closed, then for $\delta < \delta^0$ the conditions are such that the above theorem is applicable. The analytic form of the condition of separability also leads to sufficient conditions (see §5) for existence of an admissible control, in which the functional being minimized takes on a given finite value.

A complete investigation of Problem (c) has been done by S. V. Churakova.

3.2. Existence theorem for a supporting plane to a convex surface.

THEOREM 2. Let $\gamma = \gamma(w)$ be a bounded function, convex in the region Δ , $\Delta \subset W$. Then at each point $(w, \gamma(w))$ we can construct a supporting plane.

As an example of the application of this theorem, we consider the following problem for (1).

Given the numbers t_0 , T , $T > t_0$, Δ_1 , Δ_2 , and points a_1 , a_2 in X , find a control $u(t)$, $u \in U$, which minimizes the functional $\phi(x, u)$, convex on the set of elements (x, u) , where the inequalities

$$\|x(t_0) - a_1\| \leq \Delta_1, \quad \|x(T) - a_2\| \leq \Delta_2$$

must be satisfied.

In this case the set

$$\Omega(z) = \{z: z = (x, y), x = Su + c, u \in U, y = \phi(x, u)\}$$

need not be convex, but the surface $\delta(x) = \min_{x=Su+c, u \in U} \phi(x, u)$ is convex. An analogous situation occurs in control problems with certain restrictions on the function $u(t)$. The latter problems are dealt with in [14].

3.3. The L -problem in an abstract normed linear space. Given a normed linear space X of functions $h(\tau)$, $t_1 \leq \tau \leq t_2$, functions $h_i(\tau)$ from X , $i = 1, \dots, n$, and numbers c_i , L , $L > 0$, find a linear functional f over X such that

$$f(h_i) = c_i, \quad \|f\| \leq L.$$

Conditions for solvability of this problem were obtained by M. G. Krein [7]. N. N. Krasovskii [15] was the first to use the L -problem for solving a problem of time-optimal control with fixed end conditions. The class of controls was determined by the condition $\|u\| \leq L$, where the norm for $u(t)$ is given by one of the standard relations:

$$(5) \quad \|u\| = \operatorname{ess} \max_{\tau} |u_j(\tau)|, \quad t_1 \leq \tau \leq t_2,$$

$$(6) \quad \|u\| = \operatorname{ess} \max_{\tau} \left(\sum_{i=1}^n u_i^2(\tau) \right)^{1/2},$$

$$(7) \quad \|u\| = \max_j \left(\int_{t_1}^{t_2} |u_j(\tau)|^p d\tau \right)^{1/p}, \quad p \geq 1.$$

We let the symbol $[Q]_j$ denote the j th row of the matrix Q . The coordinate form of (3) leads to the equalities

$$x_i(T) - c_i = \int_{t_0}^T ([F(T, \tau)C(\tau)]_i, u(\tau)) d\tau,$$

which can be treated as the values $x_i(T) - c_i$ of a linear functional generated by the function $u(\tau)$, $t_0 \leq \tau \leq T$, with bounded norm, $\|u\| \leq L$, on the elements $[F(T, \tau)C(\tau)]_i$. This is also a typical formulation of the L -problem.

Using this approach, the authors of [16]–[18] studied limit passages from

the solution of time-optimal problems under restrictions (6), (7) as $p \rightarrow \infty$, $[C^*]_i \rightarrow 0$, $i > 2$, to the solutions of the analogous problems subject to the restrictions (5).

Subsequently the scope of problems which could be solved with the help of the L -problem was extended. The basic approach was construction of special spaces in which the norms corresponded to the type of restrictions given. Thus discrete systems with cyclic restrictions on the controlling functions were investigated in [19]. One such problem can be stated as follows.

Given an integer π , $\pi > 0$, determine N from the condition $N\pi \leq K \leq (N + 1)\pi$, where K is the duration of the process. Let $x(n + 1) = Ax(n) + Bu(n)$, $x(0) = x_0$, where n is discrete time, and A , B are constant matrices. For controls from the class

$$\max_{0 \leq s \leq N} \sum_{i=s\pi}^{(s+1)\pi-1} |u_j(i)| \leq 1,$$

we wish to determine the minimal K such that $x(K) = 0$.

The investigation of I. A. Litovchenko [20], who considers, from a functional analysis viewpoint, optimal processes with stepwise restrictions on controlling influences, is closely related to the cited work of R. Gabasov. The latter problems allow an interesting physical interpretation.

The idea of the solution of time-optimal problems with the help of the L -problem is generalized to more complicated problems in [21]. Many other problems can be reduced to the L -problem if certain transformations are made of relation (3). R. Gabasov and the author applied such reductions to problems with bounded phase coordinates, to systems connected by controls, and to systems with inertial regulators. The same approach was used in [22] in minimizing mean square error of a system.

3.4. Problem of imbeddability of convex bodies. This type is represented by the following problem: Given a normed linear space X of functions $h(\tau)$, $t_1 \leq \tau \leq t_2$, functions $h_i(\tau)$ from X , $i = 1, \dots, n$, and numbers L , Δ , $L > 0$, $\Delta > 0$, find a linear functional f over X and an element c in n -dimensional space such that

$$f(h_i) = c_i, \quad \|c\| \geq \Delta, \quad \|f\| \leq L.$$

The present situation arises, for example, when a point x moving along the trajectory of (1) is to be transferred from a given convex region to the boundary of another, also convex, region containing the first [12].

3.5. Reduction of a variational problem to a game. We consider one such problem.

Let $x(s) = Su + c(s)$, where s is a parameter, $s \in \theta$, $u \in U$, U is a normed linear space, $\|u\| \leq L$, L is a positive constant, S is a linear trans-

formation from U to W , $x(s)$, $c(s)$ for fixed s are elements of the finite-dimensional space W .

Consider the problem of minimizing the function $f(x(\cdot))$, $x(\cdot) = \{x(s), s \in \theta\}$. Letting $x(\cdot) = c(\cdot) - y$, we restrict consideration to the case where $f(c(\cdot) - y)$ is quasi-convex in y . If

$$\min_{w \in W} f(c(\cdot) - w) = f(c(\cdot) - e) = d$$

and $\max_{\|y\| \leq 1} \{(g, y) - L \|S^*g\|\} = \Delta(e) \geq 0$, then the point $y = e$ does not

belong [12] to the interior of the region $\Delta(y) = \{y: y = -Su, \|u\| \leq L\}$. Therefore the minimum δ^0 for $f(c(\cdot) - y)$ is attained on the boundary of the region $\Delta(y)$.

Let $\{g^0, y^0\}$ be a saddle point [23] of the following game:

$$(8) \quad \min_y \max_g \{f(c(\cdot) - y) + (g, y) - L \|S^*g\|\} \\ = \max_g \min_y \{f(c(\cdot) - y) + (g, y) - L \|S^*g\|\} = \alpha^0.$$

Then $(g^0, y^0) - L \|S^*g^0\| = \max_g \{(g, y^0) - L \|S^*g\|\} = 0$. Therefore $f(c(\cdot) - y^0) \leq f(c(\cdot) - y) + \max_g \{(g, y) - L \|S^*g\|\}$.

If $\min_{y=-Su, \|u\| \leq L} f(c(\cdot) - y) = f(c(\cdot) - \tilde{y})$, clearly $\tilde{y} = y^0$, $\alpha^0 = \delta^0$.

Thus the problem of minimizing the quasi-convex function $f = f(c(\cdot) - y)$, $y + Su = 0$, $\|u\| \leq L$, reduces to the game (8). A more detailed description of control problems which reduce to games will be given in §6.

4. Problem of controllability of linear systems. An important problem in the theory of optimal processes and its applications is the problem of finding a control $u(x_0, x_1, t)$ which guarantees passage of a system from the initial state x_0 to a given state x_1 . From the functional analysis point of view this two-point problem for systems (1), (2) can be interpreted as a problem of finding some linear functional (operator). The latter approach makes it possible to obtain effective conditions which guarantee existence of a control $u(x_0, x_1, t)$.

We give some definitions (cf. [24]). Let Z be the space of states of a dynamic system, U the set of control functions, $z = z(z_0, u, t)$ the state of the system at time t associated with the initial condition z_0 , $z_0 \in Z$, $z_0 = z|_{t=t_0}$ and the control u , $u \in U$. Let X denote a subspace of Z , and $x = x(z_0, u, t)$ denote the projection of the state z on X . Let θ be the zero element in Z .

DEFINITION 5. The state z_0 is called *controlled* in the class U (*controlled state*) if there exist a control¹ $u = u_{z_0}$, $u \in U$, and a number T , $t_0 \leq T < +\infty$, such that $z(z_0, u, T) = \theta$.

DEFINITION 6. The state z_0 is called *controlled* in the class U *with respect to a given set X* (*with respect to a controlled state*) if there exist a control $u = u_{z_0}$, $u \in U$, and a number T , $t_0 \leq T < +\infty$, such that $x(z_0, u, T) = \theta_x$ (θ_x is the projection of θ on X).

DEFINITION 7. If each state z_0 , $z_0 \in Z$, of a dynamic system is controlled, then we say that the system is *completely controlled*. By a *relatively controlled system* we mean a dynamic system each state z_0 of which is relatively controlled.

Consider (1) where $f(t) \equiv 0$, $x(t_0) = x_0$. Assume that in Definitions 6 and 7 the subspace X is n -dimensional. Clearly the concept of "relatively controlled state" is equivalent to the known [24] term "controlled state." The properties of completely controlled systems can be obtained from the following considerations.

Let L be a positive constant. For (1) we find $\delta^0 = \min \|x(T)\|$, $\|u\| \leq L$. From the definition of norm and the minimax theorem [23] we have:

$$\delta^0 = \min_{\|u\| \leq L} \|x(T)\| = \min_{\|u\| \leq L} \max_g \frac{(g, Su + c)}{\|g\|} = \max_g \min_{\|u\| \leq L} \frac{(g, Su + c)}{\|g\|}.$$

Thus

$$(9) \quad \delta^0 = \max_{\|g\| \leq 1} \{(g, c_1) - L \|S^*g\|\}, \quad c_1 = F(T, t_0)x_0.$$

The assertion follows.

LEMMA 1. *In order for system (1) to be completely controlled, it is necessary and sufficient that $\|S^*g\| \neq 0$ for arbitrary $g \in X^*$, $\|g\| \neq 0$.*

Necessity. Suppose system (1) is completely controlled, but there exists a vector g^0 , $\|g^0\| \neq 0$, such that $\|S^*g^0\| = 0$. Consider the set $\omega(x_0) = \{x_0 : (g^0, c_1) > 0\}$. If $x_0' \in \omega(x_0)$, then

$$\max_{\|g\| \leq 1} \{(g, F(T, t_0)x_0') - L \|S^*g\|\} \geq (g^0, F(T, t_0)x_0'), \quad L > 0.$$

Thus $\delta^0 > 0$ for all $L > 0$, which contradicts the hypothesis.

Sufficiency. Consider (9). For each x_0 the function δ^0 is continuous in L and, for sufficiently large L , negative. Therefore for some $L = L(x_0)$ the quantity δ^0 is equal to zero.

The assertion is proved.

Thus system (1) is completely controlled only if $\|S^*g\| \neq 0$, $g \in X^*$,

¹ Below we let $U = \{u : \|u\| \leq L, L < +\infty\}$.

$\|g\| \neq 0$. In the case of stationary systems,

$$\frac{dx}{dt} = Ax + Cu, \quad x = (x_1, \dots, x_n),$$

this condition turns into the requirement of linear independence of the vectors $C, AC, \dots, A^{n-1}C$, where C is the vector obtained by R. V. Gamkrelidze [25]. If C is a matrix, then, as was shown by J. P. Lasalle [26], we must require that

$$(10) \quad \text{rank} \{C, AC, \dots, A^{n-1}C\} = n.$$

Effective conditions for controllability of nonstationary systems can be obtained from Lemma 1.

Now consider the equation with delay (2). The space of states for (2) is the set of vector-valued functions

$$(11) \quad \{x(\tau), t - h \leq \tau < t\}.$$

The initial state z_0 of system (2) is determined by the conditions

$$(12) \quad z_0 = \{x_0(\tau), x_0(\tau) \equiv \phi(\tau), t_0 - h \leq \tau < t_0, x(t_0) = x_0\}.$$

The space of vectors x is a subspace of Z . The state $z = z(z_0, u, t)$ of system (2) in the space Z at time t is determined by the segment of the trajectory (11) from the space X .

Below we assume that motions of system (2) take place ($t \geq t_0$) in the space of continuous functions; A, B, C are constant matrices, U is the set of piecewise continuous functions, and $t_0 = 0$.

According to Definitions 5-7, the state (12) of system (2) is controlled if there exists a control $u, u \in U$, such that $x(t) \equiv 0, T - h \leq t \leq T$ for $T < +\infty$.

The state (12) of system (2) is controlled with respect to X if there exists a control $u, u \in U$, such that $x(T) = 0$ for $T < +\infty$.

It follows from (4), (9) that system (2) is controlled with respect to X if and only if $\|S_1^*g\| \neq 0$ for $g \in X^*, \|g\| \neq 0$ (analogue of Lemma 1).

Effective necessary and sufficient conditions for relative controllability are available for this situation and can be stated as the following theorem.

THEOREM 3. *In order for system (2) to be relatively controlled, it is necessary that the rank of the matrix*

$$(13) \quad \{P_1^1, P_1^2, P_2^2, \dots, P_{2^{n-1}}^n\},$$

where $P_1^1 = C, P_{2^l-1}^{k+1} = AP_i^k, P_{2^l}^{k+1} = BP_i^k, l = 1, \dots, 2^{k-1}, k = 1, \dots, n-1$, be equal to n .

THEOREM 4. *In order for system (2) to be relatively controlled, it is neces-*

sary and sufficient that the rank of the matrix

$$(14) \quad \{Q_1^1, Q_1^2, Q_2^2, \dots, Q_n^n\},$$

where $Q_1^1 = C$, $Q_l^{s+1} = BQ_{l-1}^s + AQ_l^s$, $l = 1, \dots, k$, $k = 0, 1, \dots, n-1$, $Q_l^s = 0$ for $l = 0$ and $l > k$, be equal to n .

In the case of differential equations without delay, sequences (13) and (14) coincide, and the conditions of Theorems 3, 4 reduce to condition (10).

The results relating to controllability described above often make it possible to study completely the problem of controllability of system (2). Consider, for example, the equation

$$(15) \quad \frac{dx}{dt} = Bx(t-h) + Cu(t).$$

The following assertion is true.

THEOREM 5. *In order for system (15) to be completely controlled it is necessary and sufficient that it be relatively controlled.*

System (2) is completely controlled if the matrix C is nonsingular. The problem is solved analogously for the equation

$$x^{(n)} + \sum_{i=1}^n (a_i x^{(n-i)}(t) + b_i x^{(n-i)}(t-h)) = cu(t),$$

which is always completely controlled.

5. Examples of reduction of control problems to functional problems. Basic relations in optimal systems. Problems of functional analysis arise in constructing admissible controls satisfying some boundary conditions (without the requirement of optimality with respect to a definite criterion). We consider several of them, with the goal of obtaining necessary and sufficient conditions for existence of solutions. We shall first show that investigation of existence and uniqueness reduces to the study of finite-dimensional extremal problems dual to those of [27].

The problem of minimizing the norm of the final state of trajectories of (1). Suppose we are given points a_1, a_2 , $a_1 \in X$, $a_2 \in X$. For given t_0 , $T > t_0$, $\Delta_1 > 0$, we wish to find a control $u^0(t)$, $\|u^0\| \leq 1$, such that

$$\|x(t_0) - a_1\| \leq \Delta_1, \quad \|x(u^0, T, t_0) - a_2\| = \min_{\|u\| \leq 1} \|x(u, T, t_0) - a_2\| = \Delta_2^0.$$

We choose a number Δ_2 and find conditions for existence of an admissible control for which

$$(16) \quad \|x(t_0) - a_1\| \leq \Delta_1, \quad \|x(T) - a_2\| \leq \Delta_2.$$

This is a functional analysis problem. Conditions for its solvability can be obtained by using the theorem on separability of convex closed sets (see §3). We give them in the form of a theorem.

THEOREM 6. *In order for problem (16) to have a solution, it is necessary and sufficient that the following inequality be satisfied:*

$$(17) \quad \begin{aligned} \max_{\|g\|=1} \Lambda(g, \Delta_2) &= \max_{\|g\|=1} \{ (g, F(T, t_0)a_1 + c_2 - a_2) \\ &\quad - \Delta_1 \|F^*(T, t_0)g\| - \Delta_2 \|g\| - \|S^*g\| \} \leq 0, \\ c_2 &= \int_{t_0}^T F(T, \tau)f(\tau) d\tau. \end{aligned}$$

The finding of optimal controls appears as the following step in the approach which uses methods of functional analysis.

In the present case we proceed as follows. The function $\Lambda(g, \Delta_2)$ is strictly monotone (decreasing) in Δ_2 ; therefore,

$$(18) \quad \Delta_2^0 = \max_{\|g\| \leq 1} \{ (g, F(T, t_0)a_1 + c_2 - a_2) - \Delta_1 \|F^*(T, t_0)g\| - \|S^*g\| \}.$$

Thus to determine Δ_2^0 we must solve the problem (18). If g^0 is a vector furnishing the solution (18), then, as follows from (17), the optimal control u^0 satisfies the condition

$$(19) \quad (S^*g^0, u^0) = \min_{\|u\| \leq 1} (S^*g^0, u).$$

Thus Theorem 6 contains necessary and sufficient conditions for existence of a solution to control problem (16), the maximum principle (19) and relation (18), making it possible to find the vector g^0 (initial condition ψ_0 for the equation conjugate to (1) for $u \equiv 0$ equal to $\{-F^*(\tau)g^0\}$). The analogous conclusion can be drawn in the following problems.

Problem of minimizing the mean square error of system (1). Suppose we want to minimize

$$J(u) = \int_{t_0}^T \left(\sum_{i=1}^n \alpha_i x_i^2(\tau) + \sum_{j=1}^r \beta_j u_j^2(\tau) \right) d\tau, \quad \alpha_i \geq 0, \quad \beta_j \geq 0, \quad \|u\| \leq 1,$$

subject to (16), where α_i, β_j are given.

As above, we begin by considering the problem

$$(20) \quad J(u) \leq \delta, \quad \|x(t_0) - a_1\| \leq \Delta_1, \quad \|x(T) - a_2\| \leq \Delta_2,$$

where conditions for solvability can be obtained by using the approach of the first theorems of §3.

THEOREM 7. *In order for problem (20) to have a solution, it is necessary*

and sufficient that²

$$\max_{f>0, \|\sigma\|=1} \left\{ (g, F(T, t_0)a_1 + c_2 - a_2) - f\delta - \Delta_1 \|F^*(T, t_0)g\| - \Delta_2 \|g\| + \min_{\|u\|\leq 1} [(S^*g, u) + fJ(u)] \right\} \leq 0.$$

For the problem of optimal control we have the following result:

$$J(u^0) = \max_{\|\sigma\|\leq 1} \left\{ (g, F(T, t_0)a_1 + c_2 - a_2) - \Delta_1 \|F^*(T, t_0)g\| - \Delta_2 \|g\| + \min_{\|u\|\leq 1} [(S^*g, u) + J(u)] \right\},$$

where the function u^0 is determined from the condition

$$(21) \quad \min_{\|u\|\leq 1} [(S^*g^0, u) + J(u)] = (S^*g^0, u^0) + J(u^0).$$

It is clear that we can pass from (17) and (20) to problems of time-optimal control subject to conditions (16) and (20), respectively. Here the optimal time of the process is the smallest number satisfying the inequality in the conditions for solvability.

We go on to a discussion of problems related to the L -problem. The latter has been used for a long time in the study of time-optimal control, in the following formulation:

Given points $x(t_0)$, $x_1 = 0$, transfer the trajectory (1) from the point $x(t_0)$ to $x_1 = 0$ in the least possible time, subject to $\|u\| \leq 1$.

Suppose the norm of $u(t)$ is given by (5). First consider the problem of transferring from the point $x(t_0)$ to the point $x_1 = 0$. Its analytic form is

$$(22) \quad -x(t_0) - \int_{t_0}^T F^{-1}(\tau)f(\tau) d\tau = \xi = \int_{t_0}^T F^{-1}(\tau)C(\tau)u(\tau) d\tau.$$

As is known [7], problem (22) has a solution if and only if

$$(23) \quad \Lambda(T) = \min_{(g, \xi)=-1} \Lambda(g, T) = \min_{(g, \xi)=-1} \int_{t_0}^T \sum_{j=1}^r |\langle g, [C^*(\tau)(F^{-1}(\tau))^*]_j \rangle| d\tau \geq 1.$$

Suppose system (1) is completely controlled. For homogeneous systems the function $\Lambda(T)$ is continuous and strictly increasing [15] in T , and therefore the least $T = T^0$ is found from the condition

$$T^0 = \max \{T: \Lambda(g, T) = 1\}.$$

In other cases (inhomogeneous system, $x_1 \neq 0$) the solution $T = T^0$ gives the smallest root of the equation $\Lambda(T) = 1$. The optimal control satisfies

² $J(u)$ is computed from (3).

the condition

$$(24) \quad \int_{t_0}^T (C^*(\tau)(F^{-1}(\tau))^* g^0, u^0(\tau)) d\tau = -1,$$

where g^0 is the solution of problem (23) for $T = T^0$.

Now consider the following problem. Given numbers $T, \Delta, \sigma, \sigma > \Delta$, find an equation for $u, \|u\| \leq 1$, such that

$$(25) \quad \|x(t_0)\| \leq \Delta, \quad \|x(T)\| \geq \sigma.$$

The solution is carried out according to the scheme of §3.4.

THEOREM 8. *In order for problem (25) to have a solution, it is necessary and sufficient that the following inequality be satisfied:*

$$(26) \quad \min_{\|g\|=1} \{\sigma \|g\| - \Delta \|F^*(T, t_0)g\| - \|S^*g\| - (g, c_2)\} \leq 0.$$

Problem (25) is related to the following problem of optimal processes.

Problem of maximizing the norm of the final state of the trajectories of (1).

Suppose we are given the numbers Δ and $\|x(t_0)\| \leq \Delta$. We wish to choose the control $u^0, \|u^0\| \leq 1$, such that

$$\|x(u^0, T, t_0)\| = \max_{\|u\| \leq 1} \|x(u, T, t_0)\| = \sigma^0.$$

From (26) we have

$$\sigma^0 = \max_{\|g\| \leq 1} \{(g, c_2) + \Delta_1 \|F^*(T, t_0)g\| + \|S^*g\|\}.$$

From the solution g^0 of this problem we determine the optimal control, since

$$(27) \quad (S^*g^0, u^0) = \max_{\|u\| \leq 1} (S^*g^0, u).$$

We emphasize once more that Theorems 6, 7, 8, and (23) contain the maximum principle (see (19), (21), (24), (27)), existence theorems for admissible controls satisfying certain boundary conditions, existence theorems for optimal controls, and conditions (18), (23), etc., for determining the quantity g^0 . These theorems also make it possible to find conditions under which the solutions of optimal control problems are continuous in the initial data and the parameters.

Remark 1. Up to now we have been concerned with controls constrained by conditions (5), (6), (7). Nonsymmetric restrictions on controls $u_j(\tau)$ of the form

$$d_j^{(1)} \leq u_j(\tau) \leq d_j^{(2)}, \quad d_j^{(1)}, d_j^{(2)} \text{ const.},$$

can be investigated by introducing a nonsymmetric norm.

Remark 2. The results obtained allow a natural passage to a discrete model. However, in connection with the fact that the corresponding functions $\Delta(T)$ (in time-optimal problems) change their values by jumps, in an investigation of existence problems uniqueness of controls gives rise to peculiarities, which were noted and studied by R. Gabasov [28].

6. Some statistical problems of optimal control. Now we discuss a stochastic model of a controlled process, considering the effect of random factors of various kinds with known probabilistic characteristics.

Suppose the random vector \tilde{x} of phase coordinates at a fixed moment of time $t = T$ has the form

$$(28) \quad \tilde{x} = \tilde{x}(u, T, t_0) = Su + \tilde{c}, \quad \tilde{x}, \tilde{c} \in X.$$

Here S is a linear operator, \tilde{c} is a random vector—the value of some operator S_1 given on the space of initial conditions, external perturbations and characteristics of another process y . For example, if \tilde{x} satisfies (1), then

$$\tilde{c} = F(T, t_0)x(t_0) + \int_{t_0}^T F(T, \tau)f(\tau) d\tau.$$

Let $f(z)$ be a positive function of z , and $Mf(\tilde{x})$ be the mathematical expectation of $f(\tilde{x})$.

DEFINITION 8. We say that the vector e , $e \in X$, is the *average* [29] *value* (average) of the random vector \tilde{c} , and the number d is the *measure of dispersion* (dispersion) if the following relation is satisfied:

$$Mf(\tilde{c} - e) = \min_z Mf(\tilde{c} - z) = d.$$

We restrict consideration to the case where $f(\tilde{x}) = \|\tilde{x}\|$. Let $\Phi_{\tilde{c}}(s)$, $s \in X$, be the distribution function of \tilde{c} .

Problem A. Find a control u^0 , $\|u^0\| \leq 1$, achieving a minimum for the functional

$$M\|\tilde{x}\| = \int_X \|\tilde{x}\| d\Phi_{\tilde{c}}(s).$$

On the basis of the results of §3.5, in the general case we have

$$(29) \quad \min_{\|u\| \leq 1} M\|\tilde{x}\| = \delta^0 = \max_g \min_z \{M\|\tilde{c} - z\| + (g, z) - \|S^*g\|\}.$$

The optimal control is found from the condition

$$(S^*g^0, u^0) = \min_{\|u\| \leq 1} (S^*g^0, u),$$

where g^0 is an element of the saddle point $\{g^0, z^0\}$ of the game (29).

Thus the problem of minimizing the mathematical expectation of the

norm of the finite state (28) reduces to the solution of the game (29). It can be shown that for the minimal value of the functional we have the estimate

$$d \leq \delta^0 \leq d + \Delta(e),$$

where e is the average of the vector \tilde{c} , $\Delta(e) = \max_{\|g\| \leq 1} \{ (g, e) - \|S^*g\| \}$, d is dispersion. Here for completely controlled systems $(g^0, z^0) > 0$, $(g^0, e) > 0$. Thus, if the object is one-dimensional (more precisely, if we are minimizing the mathematical expectation of the absolute value of one coordinate), then the optimal control is completely determined by the average e .

In certain cases Problem A is considerably simplified. Suppose, for example,

$$f(\tilde{x}) = \left(\int_X (\tilde{x}, \tilde{x}) d\Phi_{\tilde{c}}(s) \right)^{1/2};$$

then

$$(\delta^0)^2 = d^2 + \Delta^2(e), \quad e = M\tilde{c}.$$

Thus in the latter case the optimal control is completely characterized by the vector $M\tilde{c}$:

$$(S^*g^0, u^0) = \min_{\|u\| \leq 1} (S^*g^0, u),$$

$$(g^0, M\tilde{c}) - \|S^*g^0\| = \max_{\|g\| \leq 1} \{ (g, M\tilde{c}) - \|S^*g\| \}.$$

We now describe another problem encountered in applications, whose mathematical formulation leads to equations different from (28).

Suppose the action of control on an object may be discontinued at time t_1 with probability p_1 , at time t_2 with probability p_2 , etc. We are given a time $t = T$, a point $x = x_1$, and wish to minimize the mathematical expectation $M\|\tilde{x} - x_1\|$.

One of the possible ways to solve this problem is the following. In (1), assume the matrix $C(t)$ equals $\alpha(t)E(t)$, where $E(t)$ is a matrix and $\alpha(t)$ is a random process of the following special form: in the intervals $(t_k, t_{k+1}]$ the function $\alpha(t)$ can take on only two values 0 or 1 with probabilities which are related in an obvious fashion to the quantities p_k . For such a stochastic model we can use the results of §3.5 (see [29]).

Now suppose that in the phase space X we are given the point x_1 and a neighborhood of it: $\|x - x_1\| \leq \epsilon$. A cross-section of the random process \tilde{x} at a fixed moment gives a collection of random vectors $\tilde{x}(T)$ for which the quantity $\chi(u) = \text{Probability} \{ \|x - x_1\| \leq \epsilon \} = P\{\|x - x_1\| \leq \epsilon\}$ is defined.

Problem B. Find a control u^0 , $\|u^0\| \leq 1$, such that

$$\chi(u^0) = \max_{\|u\| \leq 1} \chi(u).$$

In the example of minimizing the mathematical expectation of the norm of the final state given above, we illustrated the way to reduce an infinite-dimensional variational problem to a game with two players whose strategies are finite-dimensional vectors. This method can also be applied (with certain restrictions) to Problem B. Here it turns out to be possible to obtain estimates for the maximally admissible probability without computing controls; the optimal control in the given statistical problem coincides with the optimal control in the deterministic problem:

$$x(T) = Su + x^0, \quad \min_{\|u\| \leq 1} \|Su + x^0\| = \|Su^0 + x^0\|,$$

where x^0 is an element of the saddle point of a certain finite-dimensional game. Problem B is clearly related to the determination of the minimal radius ϵ (for fixed β), where

$$P\{\|x - x_1\| \leq \epsilon\} \geq \beta,$$

whose solution we will not deal with.

7. Continuous dependence of solutions of optimal control problems on initial data and parameters. The discussion of these problems is based on the work of N. N. Krasovskii [16], [30] and the author [31]. Interesting properties of optimal controls, as functions of initial states x_0 and the parameter μ , arise in problems subject to the restrictions (5).

First we treat the problem of time-optimal control. The specific nature of the problem, whose solution—controls u —is a set of discontinuous functions, leads us to the following.

DEFINITION 9. The optimal solution $T^0(x_0, \mu)$, $u^0(x_0, \mu, t)$ is said to be *continuous* in the initial data x_0 and the parameter μ at the point (x_0^0, μ) if for each $\epsilon > 0$ there exists a $\delta > 0$ such that the inequalities

$$|T^0(x_0^0, \mu^0) - T(x_0, \mu)| < \epsilon,$$

$$\text{Meas}(E_j | u_j^0(x_0^0, \mu^0, t) - u_j^0(x_0, \mu, t) | \geq \sigma) < \epsilon, \quad \sigma > 0,$$

are satisfied, since only $\|x_0^0 - x_0\| + |\mu^0 - \mu| < \delta$.

Continuous dependence of the solutions $T^0(x_0, \mu)$, $u^0(x_0, \mu, t)$ on the initial data and the parameter was first proved for linear homogeneous systems [31]. The proof was based on the property of monotonicity [15] in T of the function $\Lambda(T)$ (see (23)). The function Λ may lose this property if transfer occurs not to the origin but into some fixed point $x_1 = a$ of the

space X . However in this case also we may obtain effective conditions guaranteeing continuous dependence of solutions on the initial data and the parameter. If the system is stationary, then one of these conditions (necessary and sufficient) is that the origin belong to the region $V = \{v: v = Aa + Cu, \|u\| \leq 1\}$.

Assume that the control system has the form

$$(30) \quad \frac{dx}{dt} = A(t, \mu)x + C(t)u + f(t, \mu), \quad \|u\| \leq 1,$$

where μ is a parameter, $\mu_1 \leq \mu \leq \mu_2$.

Suppose the minimal possible passage time of the trajectory (30) from the point (x_0, μ) to the point $(0, \mu)$ is $T^0(x_0, \mu)$.

THEOREM 9. *The optimal solution $T^0(x_0, \mu)$, $u^0(x_0, \mu, t)$ for system (30) is continuous in the initial data x_0 and the parameter μ if and only if for each positive number ν , we can construct a neighborhood $\Delta(x)$ of the point $(0, \mu)$, for points of which there exists a control u_x , $\|u_x\| \leq 1$, transferring points x into the point $x = 0$ in time $t \leq \nu$.*

In [17] the author established, for an optimal high-speed problem, existence of an optimal Lyapunov function (optimal time $T^0(x_0, \mu)$) which has continuous partial derivatives of any order in x_{i0} and the parameter μ . This fact made it possible to prove [31] that the function $T = T^0(x_0, \mu)$, subject to the restrictions (5), has continuous partial derivatives of any order in the coordinates x_{i0} , $\|x_0\| \neq 0$, and the parameter at any point which is not a control switching point.

Continuous dependence of solutions on initial data and a parameter was studied, and the appropriate sufficient conditions were also established for nonlinear systems [31].

We note that the regular properties of solutions of problems of time-optimal control, with respect to initial data in the entire space X of states of the system and the parameter μ , are inherent in problems with "smooth" restrictions, for example, of type (6).

Consequently for such problems of optimal high-speed the heuristic principle of R. Bellman can be considered strictly justified.

In considering optimal control problems with criteria other than high speed, in many cases (minimization concerning the final state, convex control function), due to the fact that the functionals allow an explicit representation (see, for example, (18)), properties of functionals, such as continuous dependence and differentiability in initial data and parameters, are easily proved. In more complicated problems, conditions guaranteeing continuous dependence of solutions of optimal control problems on initial data and parameters can be obtained on the basis of necessary and sufficient conditions for existence of solutions, examples of which are given in §5.

8. Problem of numerical solution. As was shown in §5, the methods of functional analysis applied to optimal control problems lead to additional conditions (compared to Euler equations or the maximum principle), which, as a rule, facilitate the problem of determining the initial condition for the adjoint system. Essentially this is the gradient method with large steps [32].

Thus, suppose we minimize the quantity $\|x(T)\| = (x(T), x(T))^{1/2}$ on the trajectories of (1), where $C(t) = b(t)$, $b(t)$ is the vector for the control $u(t)$ satisfying the condition $|u(t)| \leq 1$. We first show that the problem of finding the gradient reduces to integration of the original system for some specially chosen control.

It follows from (9) that

$$(31) \quad \min_{|u| \leq 1} \|x(T)\| = \Delta^0 = \max_{(g, c) \leq 1} \left\{ (g, c) - \int_{t_0}^T |(g, F(T, \tau)b(\tau))| d\tau \right\}.$$

We note once again that the vector g^0 solving problem (31) is related as follows to the initial condition ψ_0 of the equation $\dot{\psi}_0 = -F^*(T)g^0$ conjugate to the homogeneous one for (1).

We let $\lambda(g)$ denote the expression under the max in (31). Let g^1 be some vector, $\|g^1\| = 1$, and let u^1 be the control for which

$$\int_{t_0}^T (g^1, F(T, \tau)b u^1(\tau)) d\tau = \min_{|u| \leq 1} \int_{t_0}^T (g^1, F(T, \tau)b u(\tau)) d\tau.$$

We determine the point $x^1 = Su^1 + c$. It can be shown that

$$\text{grad} \frac{\lambda(g)}{\|g\|} \Big|_{g=g^1} = x^1 - \lambda(g^1)g^1, \quad \lambda(g^1) = (g^1, x^1).$$

Thus the problem of determining the gradient at each step reduces to finding x^1 and consequently to integrating (1) for $u = u^1$. This operation takes considerable time in solving problem (31), and therefore gradient methods with small step have little application here. Since it is possible to obtain information on the position of the maximum for $\lambda(g)$ in the direction of the gradient, we can use the method of steepest ascent. We shall discuss the latter for problem (31).

Suppose the vector g^1 satisfies the condition $(g^1, c) > 0$. We find u^1, x^1 as described above:

$$(S^*g^1, u^1) = \min_{|u| \leq 1} (S^*g^1, u), \quad x^1 = Su^1 + c.$$

If ϵ is a given number characterizing the accuracy of computing δ^0 , then for $\|x^1\| - (g^1, x^1) = \epsilon^1 > \epsilon$ we proceed as follows:

Assume that the process is at the k th step. Let $g^k = g^{k-1}(\alpha^{k-1})$ and find u^k, x^k such that $(S^*g^k, u^k) = \min_{|u| \leq 1} (S^*g^k, u)$, $x^k = Su^k + c$. We

introduce the element $g^k(\alpha) = (1 - \alpha)g^k + \alpha x^k$ and construct the function $\mu^k(\alpha) = \lambda(g^k(\alpha)) / \|g^k(\alpha)\|$. Let $\bar{g}^k = x^k / \|x^k\|$ and $\min_{|u| \leq 1} (S^* \bar{g}^k, u) = (S^* \bar{g}^k, u^k)$, $\bar{x}^k = S \bar{u}^k + c$. Compute

$$\mu^k(0), \quad \mu^k(1), \quad \left. \frac{d\mu^k}{d\alpha} \right|_{\alpha=0}, \quad \left. \frac{d\mu^k}{d\alpha} \right|_{\alpha=1},$$

so that then we approximate $\mu(\alpha)$ by another function. We have

$$\mu^k(0) = (g^k, x^k), \quad \mu^k(1) = (\bar{x}^k, x^k) / \|x^k\|,$$

$$\left. \frac{d\mu^k}{d\alpha} \right|_{\alpha=0} = \|x^k\|^2 - (x^k, g^k), \quad \left. \frac{d\mu^k}{d\alpha} \right|_{\alpha=1} = \left(\bar{x}^k, \frac{(g^k, x^k)x^k - (x^k, x^k)g^k}{(x^k, x^k)^{3/2}} \right).$$

Assume that $\beta(\alpha)$ is the approximating function. We compute α^k from the condition $\beta(\alpha^k) = \max_{0 \leq \alpha \leq 1} \beta(\alpha)$. If $\|x^{k+1}\| - (g^{k+1}, x^{k+1}) = \epsilon^{k+1} \leq \epsilon$, then the process stops. We note that these operations are sufficient for computing second derivatives for $\mu(\alpha)$ at the points $\alpha = 1, \alpha = 0$.

The rate of convergence of the proposed algorithm essentially depends on the method of introducing the parameter α , and on the coordinate system (λ, α) .

Successive approximation methods for other problems are described in [14], [33].

As was shown in §5, §6, with the approach described the variational (infinite-dimensional) problem reduces to operations with convex (concave), convex-concave functions of a finite number of variables. Here one also sees the clear connection of the theory of optimal processes with nonlinear convex programming [34], [35]. Problems (18), (23) and others from §5, §6 involve convex programming, and numerical algorithms of the latter can be used to construct optimal controls.

9. Application of functional analysis to problems of pursuit. We consider only problems of programmed pursuit and discuss the possibility of constructing strategies of a pursuing point [36]–[38].

Suppose that two points, x and y , are moving in n -dimensional phase space, y being pursued by x . The controls u, v , represented by points, are subject to the conditions

$$\|u\| \leq l, \quad \|v\| \leq m, \quad l, m \text{ const.} > 0.$$

The equations of motion have the form

$$(32) \quad \begin{aligned} \dot{x} &= A(t)x + C(t)u + f^1(t), & x(t_0) &= x_0, \\ \dot{y} &= E(t)y + D(t)v + f^2(t), & y(t_0) &= y_0, \end{aligned}$$

where $A(t), C(t), D(t), E(t)$ are known matrices and $f_i^1(t), f_i^2(t)$ are given functions.

Problem A. Find controls $u^0 = u^0(t_0, t)$, $v^0 = v^0(t_0, t)$ such that

$$\max_{\|v\| \leq m} \min_{\|u\| \leq l} T_\epsilon(u, v) = T_\epsilon(u^0, v^0) = T^0,$$

where $T_\epsilon(u, v)$ is the time required for the point x with control u to reach an ϵ -neighborhood of the point y , using the control v .

Problem B. Given the instants of time $t_0, t = T$, choose $u^1 = u^1(t_0, t)$, $v^1 = v^1(t_0, t)$ such that

$$\max_{\|v\| \leq m} \min_{\|u\| \leq l} \|x(u, T, t_0) - y(v, T, t_0)\| = \|x(u^1, T, t_0) - y(v^1, T, t_0)\|.$$

Equations (32) are linear, and therefore, under the condition that motion is considered from the instant $t = t_0$, we have the representation

$$z = Su + Qv + c$$

for the vector $z = x - y$ at the moment $t = T$. Here $S = S(T, t_0)$, $Q = Q(T, t_0)$ are linear operators and $c = c(T, t_0)$ is a vector (cf. (3)).

It follows from Theorem 4 that the point x reaches an ϵ -neighborhood of the point y , using the control v , only when

$$(33) \quad \lambda(v, T, t_0) = \max_{\|g\|=1} \{(g, c + Qv) - \epsilon \|g\| - l \|S^*g\|\} \leq 0.$$

The least $T = T(v)$ satisfying (33) is equal to the minimal time of pursuit of the point y with control v : $T(v) = \min_{\|u\| \leq l} T_\epsilon(u, v)$. Therefore for T^0 we have

$$T^0 = \max_{\|v\| \leq m} T(v) = T(v^0).$$

The control v^0 , substituted into (33), determines u^0 : $(S^*g^0, u^0) = \min_{\|u\| \leq l} (S^*g^0, u)$, where g^0 is the solution of problem (33) with $v = v^0$, $T = T^0$.

We introduce the functions

$$(34) \quad \begin{aligned} \Lambda(T, t_0) &= \max_{\|v\| \leq m} \lambda(v, T, t_0) \\ &= \max_{\|g\|=1} \{(g, c) - \epsilon \|g\| - l \|S^*g\| + m \|Q^*g\|\}, \end{aligned}$$

$$\Lambda^+(T, t_0) = \begin{cases} \Lambda(T, t_0) & \text{for } \Lambda(T, t_0) > 0, \\ 0 & \text{for } \Lambda(T, t_0) \leq 0. \end{cases}$$

Let $\lambda = \lambda(v, T, t_0)$, $\|v\| \leq m$, be continuous from the right in T .

The smallest of the numbers θ satisfying the condition $\Lambda^+(\theta, t_0) = \min_{T \geq t_0} \Lambda^+(T, t_0)$ is denoted by T^θ , and the vector g solving (34) for $T = T^\theta$ is denoted by g^θ .

DEFINITION 10. The functions $u^\theta(t_0, s)$, $v^\theta(t_0, s)$, $t_0 \leq s \leq T^\theta$, con-

structured using the relations

$$(S^*g^\theta, u^\theta) = \min_{\|u\| \leq l} (S^*g^\theta, u), \quad (Q^*g^\theta, v^\theta) = \max_{\|v\| \leq m} (Q^*g^\theta, v),$$

are called θ -optimal controls.

It is easier to find θ -optimal controls in the sense defined than to determine the functions $u^0(t_0, t)$, $v^0(t_0, t)$. Therefore it is of interest to find the relationship between the numbers T^0 , T^θ and the controls u^0 , u^θ , v^0 , v^θ .

From the definition of the numbers T^0 , T^θ it follows that $T^\theta \geq T^0$. Let $\Lambda(T^\theta, t_0) = 0$.

THEOREM 10. *If $\lambda(v^\theta, T, t_0) > 0$, $t_0 \leq T < T^\theta$, then*

$$T^\theta = T^0, \quad u^\theta(t_0, s) = u^0(t_0, s), \quad v^\theta(t_0, s) = v^0(t_0, s), \quad t_0 \leq s \leq T^0.$$

If $T = \theta^1$ is the smallest number for which $\lambda(v^\theta, T, t_0) \leq 0$, $\theta^1 < T^\theta$, and there does not exist v , $\|v\| \leq m$, with the properties $\lambda(v, T, t_0) > 0$, $t_0 \leq T \leq \theta^1$, then $T^0 = \theta^1$, $u^{\theta^1}(t_0, s) = u^0(t_0, s)$, $v^{\theta^1}(t_0, s) = v^0(t_0, s)$, $t_0 \leq s \leq \theta^1$.

The strategy $u = u^\theta(s_0, s)$ for the point x guarantees transfer into an ϵ -neighborhood of the point y for any choice of control v in the time $t \leq T^\theta - t_0$.

If the situation is such that the point x knows the position of the point y at time t and the equation of motion (32), and according to these data a strategy is to be constructed which guarantees reaching an ϵ -neighborhood of y in the least possible time, then on the basis of the above we conclude that for $\Lambda(T^\theta, t_0) = 0$, and the conditions of Theorem 10, optimal pursuit can be achieved with θ -optimal controls.

Note that if the conditions of Theorem 10 are not satisfied, then it makes sense to release the point y from the θ -optimal control, since then $T(v^\theta) < T(v^0)$.

We proceed to Problem B. For given $v(s)$, $t_0 \leq s \leq T$, the minimal distance $\delta(v, T, t_0)$ which the point x can approach at time $s = T$ is

$$\delta(v, T, t_0) = \max_{\|g\| \leq 1} \{(g, c + Qv) - l \|S^*g\|\}$$

and

$$\max_{\|v\| \leq m} \delta(v, T, t_0) = \max_{\|g\| \leq 1} \{(g, c) - l \|S^*g\| + m \|Q^*g\|\}.$$

Thus the θ -optimal controls constructed according to the relations (33), (34) coincide with the controls $u^1(t_0, t)$, $v^1(t_0, t)$ for Problem B.

We let $u(t, s)$ denote the optimal control for Problem B with initial time t , computed at time s .

If the point x knows the technical capabilities of y (the system of equations of motion) and the position of y at each moment s , then the strategy

$u = u(t, t)$ guarantees optimal pursuit in the sense of minimal motion away from y .

Other optimal pursuit problems are the subject of a special exposition.

10. On possible generalizations. It was already mentioned in the introduction that the methods of investigation described are applicable to a wide range of problems. Above, in the presentation of the basic ideas, we chose the simplest systems. Now we describe several possible ways to generalize the results obtained.

(i) In studying optimal processes for ordinary differential equations we first introduced (3). But this relation is a general property of linear systems (partial differential, integrodifferential, integral and other equations), and one can usually arrive at it with the help of Green's functions or other analogous means. Therefore the methods of functional analysis described are also applicable here.

(ii) Also for simplicity of presentation, the class of functions was defined by the condition $\|u\| \leq 1$. This restriction can be removed, choosing as a restriction on the controls any bounded convex closed set in the space of vector-functions $u(t)$. (Such problems were stated in §3.)

(iii) Restrictions on the controls can also be weakened in another direction. Namely, instead of (1), consider the equation

$$(35) \quad \frac{dx}{dt} = A(t)x + \phi(u, t), \quad u \in U,$$

where $\phi(u, t)$ is a vector-function continuous in u, t , and U is a bounded closed set.

It is possible to extend the given results to (35) because the set of admissibility for it is convex and closed. The latter fact was proved in work by H. Halkin [39], L. W. Neustadt [40].

(iv) The reasoning by which the results described for linear systems are extended to nonlinear systems,

$$(36) \quad \frac{dx}{dt} = f(x, u, t),$$

is presented in work by N. N. Krasovskii [30] and the author [31].

Unfortunately, generalizations in this direction are less effective, since one of the advantages of the methods of functional analysis (related to the determination of g^0) disappears in this case, although it is possible to investigate qualitative problems in the theory of optimal processes for (36), [30], [31].

(5) Generalizations involving passage to the infinite-dimensional case subject to (6) (or (7), $p > 1$), as is known, are not difficult, since the

theorems of functional analysis presented in §3 remain valid in this case [5]–[7]. But the effectiveness of the solutions is decreased, since the infinite-dimensional (variational) problem again reduces to the finite-dimensional. However, in certain cases the methods of functional analysis from the above point of view may also be of interest in the infinite-dimensional case [41].

It is natural that the considerations presented above relate to deterministic and stochastic systems as well as to systems with adaptation. In the latter case the necessary computations increase extraordinarily.

REFERENCES

- [1] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [2] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.
- [3] I. A. LITOVCHENKO, *Theory of optimal systems*, Sbornik Itogi Nauki (Matematicheskii analiz, teoriya veroyatnostei, regulirovanie, 1962), Nauka, Moscow, 1964.
- [4] N. N. KRASOVSKII, *Optimal control in regular dynamical systems*, Russian Math. Surveys, 20 (1965), no. 3, pp. 153–174.
- [5] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Interscience, New York, 1958.
- [6] D. P. MIL'MAN, *Separation of nonlinear functionals and their linear extensions*, Izv. Akad. Nauk SSSR Ser. Mat., 27 (1963), pp. 1189–1210.
- [7] N. I. AHIEZER AND M. G. KREIN, *Some Questions in the Theory of Moments*, Translations of Mathematical Monographs, vol. 2, American Mathematical Society, Providence, 1962.
- [8] L. A. LYUSTERNIK AND V. I. SOBOLEV, *Elements of Functional Analysis*, Ungar, New York, 1961.
- [9] R. BELLMAN AND I. GLICKSBERG, *On the "bang-bang" control problem*, Quart. Appl. Math., 14 (1956), pp. 11–18.
- [10] H. A. ANTOSIEWICZ, *Linear control systems*, Arch. Rational Mech. Anal., 12 (1963), pp. 313–324.
- [11] R. GABASOV AND F. M. KIRILLOVA, *On application of the theory of linear inequalities to optimal control*, Annotatsii dokladov Vtorogo Vsesoyuznogo s"ezda po teoret. i prikl. mekhanike, Nauka, Moscow, 1964.
- [12] ———, *The solution of some problems in the theory of optimal processes*, Automat. Remote Control, AC-25 (1964), pp. 945–955.
- [13] FAN' TSI, *On systems of linear inequalities*, Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, 1956.
- [14] R. GABASOV AND F. M. KIRILLOVA, *Optimization of convex functionals on trajectories of linear systems*, Soviet Math. Dokl., 5 (1964), pp. 745–749.
- [15] N. N. KRASOVSKII, *On an optimal control problem*, Prikl. Mat. Meh., 21 (1957), pp. 670–677.
- [16] ———, *On the theory of optimal control*, J. Appl. Math. Mech., 23 (1959), pp. 899–919.

- [17] F. M. KIRILLOVA, *A limiting process in the solution of an optimal control problem*, Ibid., 24 (1960), pp. 398-405.
- [18] R. KULIKOVSKII, *Optimal processes and synthesis of optimum automatic control systems with nonlinear invariable elements*, Proceedings of the First IFAC Congress, Moscow, 1960 (Automatic and Remote Control, vol. 1, Butterworths, London, 1961), pp. 469-476.
- [19] R. GABASOV, *Optimal processes with cyclic constraints*, Soviet Math. Dokl., 3 (1962), pp. 787-791.
- [20] I. A. LITOVCHENKO, *System optimization with step-function limitations on the control*, Automat. Remote Control (8), 26 (1965), pp. 1293-1303.
- [21] R. GABASOV, *On optimal processes in coupled digital systems*, Ibid. (7), 23 (1962), pp. 808-817.
- [22] A. B. KURZHANSKII, *Optimal control design using a method of moments and minimizing the standard deviation*, Ibid. (5), 25 (1964), pp. 568-574.
- [23] Symposium on *Infinite Antagonistic Games*, Gosudarstvennoe izd-vo fiz. mat. literaturi, Moscow, 1963.
- [24] R. E. KALMAN, *On the general theory of control systems*, Proceedings of the First IFAC Congress, Moscow, 1960 (Automatic and Remote Control, vol. 1, Butterworths, London, 1961), pp. 481-492.
- [25] R. V. GAMKRELIDZE, *Theory of high-speed optimal processes in linear systems*, Izv. Akad. Nauk SSSR Ser. Mat., 22 (1958), pp. 449-474.
- [26] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960.
- [27] G. SH. RUBINSHTEIN, *Dual extremal problems*, Soviet Math. Dokl., 4 (1963), pp. 1309-1312.
- [28] R. GABASOV, *Some problems in the qualitative theory of controlled systems*, Doctoral dissertation, Kazan State University, 1963.
- [29] R. GABASOV AND F. M. KIRILLOVA, *Statistical problem of optimal control of a linear system*, Dokl. Akad. Nauk SSSR, 164 (1965), pp. 16-19.
- [30] N. N. KRASOVSKII, *Choice of parameters of the optimal stable systems*, Proceedings of the First IFAC Congress, Moscow, 1960 (Automatic and Remote Control, vol. 1, Butterworths, London, 1961), pp. 465-468.
- [31] F. M. KIRILLOVA, *On the correctness of the formulation of an optimal control problem*, this Journal, 1 (1963), pp. 224-239; *On continuous dependence of the solution of an optimal control problem on initial data and parameters*, Uspehi Mat. Nauk, 17 (1962), pp. 141-146; *Problems in qualitative optimal control theory*, Doctoral dissertation, Moscow State University, 1962.
- [32] L. V. KANTOROVICH, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk, 31 (1948), pp. 89-185.
- [33] R. GABASOV AND F. M. KIRILLOVA, *Construction of successive approximations for certain optimal control problems*, Automat. i Telemekh., 27 (1966), pp. 5-17.
- [34] L. W. NEUSTADT, *Optimization, a moment problem and nonlinear programming*, this Journal, 2 (1964), pp. 33-53.
- [35] K. J. ARROW, L. HURWICZ AND H. UZAWA, *Studies in Linear and Nonlinear Programming*, Stanford University Press, Stanford, 1958.
- [36] L. S. PONTRYAGIN, *On some differential games*, Soviet Math. Dokl., 5 (1964), pp. 712-716.
- [37] M. N. OĞUZTÖRELİ, *Optimal pursuit strategy processes with retarded control systems*, this Journal, 2 (1964), pp. 89-105.

- [38] N. N. KRASOVSKII, YU. M. REPIN AND V. E. TRET'YAKOV, *Theory of games in control systems*, Engineering Cybernetics, 4 (1965), pp. 1-11.
- [39] H. HALKIN, *Liapunov's theorem on the range of a vector measure and Pontryagin's maximum principle*, Arch. Rational Mech. Anal., 10 (1962), pp. 296-304.
- [40] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110-117.
- [41] A. G. BUTKOVSKII AND L. N. POLTAVSKII, *Optimal control of a distributed oscillatory system*, Automat. Remote Control, 26 (1965), pp. 1835-1848.

AN EXTENSION OF AN INFORMATION-THEORETIC DERIVATION OF CERTAIN LIMIT RELATIONS FOR A MARKOV CHAIN*

S. KULLBACK†

In [1] a limit relation for the transition probabilities of a stationary Markov chain with a countable number of states was derived by the use of certain properties of information measures. In this paper we shall use essentially the same techniques to derive a limit relation for a Markov chain with a countable number of states but with constant transition probabilities only. We shall assume that the reader is familiar with [1] and shall therefore omit certain details.

Consider a Markov chain with constant transition probabilities

$$(1) \quad P = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \cdot & \cdot & \cdots \\ p_{n1} & p_{n2} & \cdots \\ \cdot & \cdot & \cdots \end{bmatrix},$$

where

$$(2) \quad \sum_j p_{ij} = 1, \quad i = 1, 2, \cdots, \quad p_{ij} > 0,$$

with the absolute distributions

$$(3) \quad \mu_j^{(m+1)} = \sum_i \mu_i^{(m)} p_{ij}, \quad j = 1, 2, \cdots, \quad \sum_j \mu_j^{(m)} = 1,$$

and with the m -step transition probabilities

$$(4) \quad p_{jk}^{(m+1)} = \sum_h p_{jh} p_{hk}^{(m)} = \sum_h p_{jh}^{(m)} p_{hk}, \quad j, k = 1, 2, \cdots,$$

$$(5) \quad \sum_k p_{jk}^{(m)} = 1, \quad j = 1, 2, \cdots$$

We now prove the following theorem.

THEOREM. *For a Markov chain with a countable number of states and constant transition probabilities, $\lim_{m \rightarrow \infty} (p_{hk}^{(m)} / \mu_k^{(m)}) = 1$.*

Consider the discrimination information between the systems of probabilities (see [1])

$$(6) \quad P_i^{(m)} : \{p_{i1}^{(m)}, p_{i2}^{(m)}, \cdots\}, \quad U^{(m)} : \{\mu_1^{(m)}, \mu_2^{(m)}, \cdots\},$$

* Received by the editors August 10, 1966.

† Department of Statistics, George Washington University, Washington, D. C. This work was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR 932-65.

$$(7) \quad P_i^{(m+1)} : \{p_{i1}^{(m+1)}, p_{i2}^{(m+1)}, \dots\}, \quad U^{(m+1)} : \{\mu_1^{(m+1)}, \mu_2^{(m+1)}, \dots\},$$

given by

$$(8) \quad I(P_i^{(m)}; U^{(m)}) = \sum_j p_{ij}^{(m)} \log \frac{p_{ij}^{(m)}}{\mu_j^{(m)}},$$

$$(9) \quad I(P_i^{(m+1)}; U^{(m+1)}) = \sum_j p_{ij}^{(m+1)} \log \frac{p_{ij}^{(m+1)}}{\mu_j^{(m+1)}}.$$

Using the convexity property and (2), (3) and (4) (see [1]) it may be shown that

$$(10) \quad I(P_i^{(m)}; U^{(m)}) \geq I(P_i^{(m+1)}; U^{(m+1)}),$$

with equality if and only if $p_{ih}^{(m)} = \mu_h^{(m)}$. The convexity also implies that

$$(11) \quad I(P_i^{(m)}; U^{(m)}) \geq 0$$

for all m , with equality if and only if $p_{ih}^{(m)} = \mu_h^{(m)}$. Thus

$$(12) \quad \begin{aligned} I(P_i^{(1)}; U^{(1)}) &\geq I(P_i^{(2)}; U^{(2)}) \geq \dots \geq I(P_i^{(m)}; U^{(m)}) \\ &\geq I(P_i^{(m+1)}; U^{(m+1)}) \geq \dots \geq 0. \end{aligned}$$

Let us assume that $I(P_i^{(1)}; U^{(1)}) < \infty$. If in (12)

$$I(P_i^{(m)}; U^{(m)}) = I(P_i^{(m+1)}; U^{(m+1)}),$$

then $p_{ih}^{(N)}/\mu_h^{(N)} = 1$, $I(P_i^{(N)}; U^{(N)}) = 0$, $N \geq m$.

On the other hand, if the sequence in (12) is strictly monotonic, and there is no equality, then the fact that $I(P_i^{(m)}; U^{(m)})$ tends to a finite limit as $m \rightarrow \infty$, and hence $I(P_i^{(m)}; U^{(m)}) - I(P_i^{(m+1)}; U^{(m+1)}) \rightarrow 0$ as $m, n \rightarrow \infty$, implies that (see [1])

$$(13) \quad \frac{p_{ih}^{(m)}}{\mu_h^{(m)}} - \frac{p_{ij}^{(m+n)}}{\mu_j^{(m+n)}} \rightarrow 0 \quad \text{as } m, n \rightarrow \infty, \quad i, j, h, = 1, 2, \dots$$

For $h = j$ in (13) the Cauchy mutual convergence criterion implies that there is a C_{ij} such that

$$(14) \quad \lim_{m \rightarrow \infty} \frac{p_{ij}^{(m)}}{\mu_j^{(m)}} = C_{ij}.$$

Thus letting $m \rightarrow \infty$ in (13) we get

$$(15) \quad C_{ih} = C_{ij} = C_i \quad \text{for all } i, j, h.$$

From (3) and

$$(16) \quad \sum_j \frac{p_{ij}^{(m)}}{\mu_j^{(m)}} \mu_j^{(m)} = 1, \quad m = 1, 2, \dots$$

it follows that

$$(17) \quad \sup_j \frac{p_{ij}^{(m)}}{\mu_j^{(m)}} \geq 1, \quad \inf_j \frac{p_{ij}^{(m)}}{\mu_j^{(m)}} \leq 1, \quad m = 1, 2, \dots.$$

Since

$$\lim_{m \rightarrow \infty} \frac{p_{ij}^{(m)}}{\mu_j^{(m)}} = C_i \quad \text{for } j = 1, 2, \dots,$$

(17) implies that $C_i \geq 1$, $C_i \leq 1$, that is, $C_i = 1$, and the theorem is proved.

Note that if $\mu_j^{(m)} = \mu_j$ for all j and m then the result here becomes that in [1].

REFERENCE

- [1] S. KULLBACK, *An information-theoretic derivation of certain limit relations for a stationary Markov chain*, this Journal, 4 (1966), pp. 454-459.

CONSTRUCTION OF SUBOPTIMAL CONTROL SEQUENCES*

R. J. LEAKE AND RUEY-WEN LIU†

Summary. As an alternative to a direct solution of the Hamilton-Jacobi equation, results are presented for the determination and improvement of suboptimal controls and for obtaining bounds on the optimal value of the performance index. Treatment is restricted to the class of problems included in the well-known work of Kalman [1].

1. Introduction. It is well established that direct attempts to solve the Hamilton-Jacobi equation in order to obtain optimal feedback controls are hopeless in all but a few special cases. Bellman foresaw these difficulties in his early work on dynamic programming [2], and in addition to his main constructive method of solution, laid strong emphasis on the use of *successive approximations* for the study of optimal processes. A similar idea for obtaining suboptimal controls based on the relationship between Hamiltonian functions and performance index derivatives has been exploited by Rekasius [3] and Haussler [4]. Their work has shown promise for the special case of "stationary" problems, with a separable scalar control. The same line of reasoning will be applied here to a broader class of problems.

Consider a dynamical system represented by

$$(1) \quad \dot{x} = f(x, k, t), \quad x(t_0) = x_0,$$

where the n -vector x is the plant state, f is a continuously differentiable n -vector function, and $k(x, t)$ is an r -vector function defined on $R^n \times R^1$. The solution of (1) will be denoted as $\phi_k(t) \triangleq \phi_k(t; x_0, t_0)$.

Let G be a closed subset of $R^n \times R^1$ to which all motions of (1) are restricted, and let the target set S be a closed subset of G . For our purposes, the function k will be called an *admissible feedback control law* if:

- (a) it is continuously differentiable with values $k(x, t)$ belonging to a locally compact set $U \subset R^r$ for all t ;
- (b) it has the property that when substituted into (1), any motion beginning in $G - S$ reaches S , or approaches S , in a uniform asymptotic manner without leaving G .

The class of functions satisfying the above properties will be denoted as \mathcal{K}^0 .

The terminal time $t_1 = t_1(x_0, t_0)$ will be defined as the first instant after t_0 when the motion $(\phi_k(t), t)$ becomes a member of S ; or, in the asymptotic case, $t_1 = \infty$.

* Received by the editors August 31, 1965, and in revised form July 19, 1966.

† Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana. This research was supported in part by the National Science Foundation under Grant GK-91.

The system performance is evaluated by the functional

$$(2) \quad J(x_0, t_0; k) = \lambda[\phi_k(t_1; x_0, t_0), t_1] \\ + \int_{t_0}^{t_1} L[\phi_k(\alpha; x_0, t_0), k(\phi_k(\alpha; x_0, t_0), \alpha), \alpha] d\alpha,$$

where L and λ are continuously differentiable functions. We define

$$V^0(x_0, t_0) \triangleq \inf_{k \in K^0} J(x_0, t_0; k).$$

Let H be defined as

$$(3) \quad H(x, p, t, u) \triangleq \langle f(x, u, t), p \rangle + L(x, u, t),$$

where p is an n -vector, u is an r -vector and $\langle \cdot, \cdot \rangle$ denotes the inner product. Assume that H has a unique absolute minimum for each x, p , and t with respect to the values $u \in U$, and let the associated location of the minimum be denoted as $c(x, p, t)$. Assuming that c is a continuously differentiable function of x, p and t , we define the Hamiltonian as

$$(4) \quad H^0(x, p, t) \triangleq H(x, p, t, c(x, p, t)) = \min_{u \in U} H(x, p, t, u)$$

and the Hamilton-Jacobi equation as

$$(5) \quad V_t + H^0(x, V_x, t) = 0,$$

where $V(x, t)$ is a scalar function defined on $R^n \times R^1$, $V_t = \partial V / \partial t$ and $V_x = \text{grad } V$. Kalman [1] has shown that if $V(x, t)$ is twice continuously differentiable in all arguments, if it satisfies (5) in G and the boundary condition $V(x, t) = \lambda(x, t)$ on S , and in addition if the function $k^0(x, t) = c(x, V_x(x, t), t)$ is admissible, then $V(x, t) = V^0(x, t)$.

The solution of (1) is easily seen to have the property that

$$(6) \quad \phi_k(\alpha; \phi_k(t; x_0, t_0), t) = \phi_k(\alpha; x_0, t_0) = \phi_k(\alpha);$$

and furthermore, if $(x_0, t_0) \in G - S$, we have for the terminal time

$$(7) \quad t_1(\phi_k(t; x_0, t_0), t) = t_1(x_0, t_0), \quad t_0 \leq t \leq t_1(x_0, t_0).$$

It follows from (2), (6) and (7) that

$$(8) \quad J(\phi_k(t; x_0, t_0), t; k) = \lambda[\phi_k(t_1; x_0, t_0), t_1] \\ + \int_t^{t_1} L[\phi_k(\alpha; x_0, t_0), k(\phi_k(\alpha; x_0, t_0), \alpha), \alpha] d\alpha$$

for $t_0 \leq t \leq t_1$. Consequently, the Eulerian derivative of J along motions

of (1) is given by

$$(9) \quad \frac{dJ}{dt}(\phi_k(t; x_0, t_0), t; k) = -L[\phi_k(t; x_0, t_0), k(\phi_k(t; x_0, t_0), t), t]$$

or, denoting $\phi_k(t; x_0, t_0)$ by x , we have in more casual notation

$$(10) \quad \dot{J}(x, t; k) = -L(x, k, t).$$

Whenever there is no chance of confusion we will use the latter representation of the result, it being understood that (9) is implied.

2. Transformations of the successive approximation. Given any optimal feedback control problem described above, define \mathcal{V} as the set of all continuously differentiable functions $V: R^n \times R^1 \rightarrow R^1$ such that $V(x, t) = \lambda(x, t)$ on S . Let \mathcal{V}^0 be the subset of \mathcal{V} such that if $k(x, t) = c(x, V_x(x, t), t)$ then $k \in \mathcal{K}^0$, i.e., k is admissible. Note that by our assumptions, if it exists, $V^0 \in \mathcal{V}^0$.

Next we define the basic transformations used to generate suboptimal control sequences.

- (a) $T_1: \mathcal{V}^0 \rightarrow \mathcal{K}^0$ is defined for any $V \in \mathcal{V}^0$ by $T_1(V) = k$, where $k(x, t) = c(x, V_x, t)$.
- (b) $T_2: \mathcal{K}^0 \rightarrow \mathcal{V}$ is defined by $T_2(k) = V$, where $V(x, t) = J(x, t; k)$. Clearly $J(x, t; k) = \lambda(x, t)$ on S , and since λ , L , and k are continuously differentiable, so is V .
- (c) $T: \mathcal{V}^0 \rightarrow \mathcal{V}$ is the composite mapping defined for $V \in \mathcal{V}^0$ by $T(V) = T_2(T_1(V)) = J(x, t; k)$ with $k(x, t) = c(x, V_x, t)$.

3. Development of the basic inequalities.

LEMMA 1. Suppose $V \in \mathcal{V}^0$ and $W \in \mathcal{V}$. Then, if

$$(11) \quad H^0(x, V_x, t) + V_t \leq H^0(x, W_x, t) + W_t, \quad (x, t) \in G - S,$$

it follows that

$$(12) \quad W(x, t) \leq V(x, t), \quad (x, t) \in G.$$

Proof. Since $V \in \mathcal{V}^0$ the control $k(x, t) = c(x, V_x, t)$ causes the motion of (1) to enter S from any initial phase in G . Taking derivatives along this motion we have

$$\begin{aligned} \dot{V} - \dot{W} &= \langle f(x, k, t), V_x \rangle - \langle f(x, k, t), W_x \rangle + V_t - W_t \\ &= [H^0(x, V_x, t) + V_t - H^0(x, W_x, t) - W_t] \\ &\quad + [H^0(x, W_x, t) - H(x, W_x, t, k)] \\ &\leq 0, \end{aligned}$$

since the term in the first bracket is nonpositive by assumption, whereas

the term in the second bracket is nonpositive by the definition of H^0 in (4). It follows that along the motion of the indicated system $\dot{V}(x, t) \leq \dot{W}(x, t)$. Noting that both Eulerian derivatives are taken along the same motion and therefore with a common termination point on S , a simple integration yields (12).

THEOREM 1. *Let $V \in \mathcal{V}^0$ and $W \in \mathcal{V}$. Then for $(x, t) \in G$,*

$$(13) \quad \begin{aligned} H^0(x, V_x, t) + V_t &\leq 0 \text{ implies } V^0(x, t) \leq V(x, t), \\ H^0(x, W_x, t) + W_t &\geq 0 \text{ implies } W(x, t) \leq V^0(x, t). \end{aligned}$$

Proof. The conclusion follows from Lemma 1 together with the fact that $V^0 \in \mathcal{V}^0$ and $H^0(x, V_x^0, t) + V_t^0 = 0$. Note that all of the above results also hold for strict inequalities. Theorem 1 provides a basic means of determining upper and lower bounds on $V^0(x, t)$.

The next result gives an alternate method of evaluating the performance index for a given admissible control law.

LEMMA 2. *Let $W \in \mathcal{V}$ and $k \in \mathcal{K}^0$. Then*

$$(14) \quad W = T_1(k) \text{ if and only if } H(x, W_x, t, k) + W_t = 0 \text{ in } G - S.$$

Proof. Assume first that $H(x, W_x, t, k) + W_t = 0$. Taking the derivative of $W(x, k)$ along motions of (1) with the control k we have

$$\dot{W} = H(x, W_x, t, k) + W_t - L(x, k, t) = -L(x, k, t);$$

but from (10), $\dot{J} = -L(x, k, t)$. Noting that the restrictions on f, k and L imply J is continuously differentiable in G and integrating along the motion of (1) yield $W(x, t) = J(x, t; k)$, i.e., $W = T_1(k)$. Conversely, $W = T_1(k)$ implies $\dot{W} = -L$ which leads to $H(x, W_x, t, k) + W_t = 0$.

Let us now turn to the composite mapping $T = T_2 T_1$.

THEOREM 2. *Let $V \in \mathcal{V}^0$ and $W \in \mathcal{V}$. Then*

$$(15) \quad W = T(V) \text{ if and only if } H(x, W_x, t, c(x, V_x, t)) + W_t = 0 \text{ in } G - S.$$

The proof follows directly from Lemma 2 and the smoothness assumption on $c(x, V_x(x, t), t) = k(x, t) = T_1(V)(x, t)$.

Notice that Lemma 2 and Theorem 2 both give methods of determining $W \in \mathcal{V}$ such that $H(x, W_x, t, k) + W_t = 0$ and thus $H^0(x, W_x, t) + W_t \leq 0$.

If it turns out that $W \in \mathcal{V}^0$, the function takes on a usefulness which is now indicated.

THEOREM 3. *Let $W \in \mathcal{V}^0$ and $W^* = T(W)$. Then*

$$(16) \quad H^0(x, W_x, t) + W_t \leq 0 \text{ in } G - S$$

implies that

$$(17) \quad V^0(x, t) \leq W^*(x, t) \leq W(x, t).$$

Proof. Taking derivatives of W and W^* along the same motion of (1) with $k(x, t) = c(x, W_x, t)$ we have

$$\dot{W} = \langle f(x, k, t), W_x \rangle + W_t = H^0(x, W_x, t) + W_t - L(x, k, t) \leq -L(x, k, t),$$

whereas from (10), $\dot{W}^* = J(x, t; k) = -L(x, k, t)$. The desired result follows by integration.

It is easy to see that V^0 is a fixed point of the mapping T , but the converse is not obvious.

THEOREM 4. *If V^0 exists, then V^0 is a fixed point of T , i.e., $V^0 = T(V^0)$. Conversely, if V is a fixed point, then $V(x, t) = V^0(x, t)$ on G .*

Proof. If V^0 exists, then $V^0 \in \mathcal{V}^0$. Let $V^* = T(V^0)$. From Theorem 3 we have $V^0(x, t) \leq V^*(x, t) \leq V^0(x, t)$, so $V^* = V^0$. Conversely, assume $V = T(V)$; then by Theorem 2, $H(x, V_x, t, k) + V_t = 0$, where $k(x, t) = c(x, V_x, t)$, i.e., $H^0(x, V_x, t) + V_t = 0$ or V is a solution of the Hamilton-Jacobi equation. With the assumption that V is a member of \mathcal{V}^0 , the domain of T , we have $V = V^0$.

Note that if V^0 exists, it is unique.

COROLLARY 1. *If V^0 exists, there is one and only one fixed point of T in \mathcal{V}^0 .*

4. Iterations and convergence conditions. The question now arises as to whether the process outlined in Theorem 3 can be continued. Given any $V^1 \in \mathcal{V}^0$ we establish the successive approximation procedure through T , i.e., $V^{n+1} = T(V^n)$, by making the following assumption.

ASSUMPTION. $T(V^n) \in \mathcal{V}^0$ for $n = 1, 2, 3, \dots$.

THEOREM 5. *If $V^1 \in \mathcal{V}^0$ and $V^{n+1} = T(V^n)$, $n = 1, 2, 3, \dots$, then*

$$V^0(x, t) \leq V^{n+1}(x, t) \leq V^n(x, t) \leq V^1(x, t), \quad (x, t) \in G.$$

Proof. Let $k^{n+1}(x, t) = c(x, V_x^n, t) = T_1(V^n)$, and note that we have defined $V^n(x, t) = J(x, t; k^n) = T_2(k^n)$. By Theorem 2 we have for each n

$$H(x, V_x^n, t, k^n) + V_t^n = 0;$$

thus $H^0(x, V_x^n, t) + V_t^n \leq 0$, and the conclusion follows inductively from Theorem 3.

Theorem 5 indicates that the estimate of $V^0(x, t)$ can only be improved by the successive approximations. We may state stronger results if further restrictions are imposed.

THEOREM 6. *If the iteration procedure terminates in a finite number of steps, that is, if $T(V^n) = V^n$ for some finite n , then V^0 exists and $V^0(x, t) = V^n(x, t)$ on G .*

Proof. Since V^n is a fixed point in \mathcal{V}^0 , the result follows directly from Theorem 4.

If the iteration does not end in a finite number of steps, Lemma 3 follows from Theorem 5.

LEMMA 3. *For every sequence $\{V^n\}$, there exists a function V^* such that $V^n(x, t) \downarrow V^*(x, t)$ pointwise on G . If G is bounded, the convergence is uniform.*

In order that $V^* = V^0$, it is required that $V^* \in \mathcal{V}^0$. If in addition T is continuous at V^* , the condition $V^* = V^0$ is assured. In order to make the statement more precise, let us assume G is bounded and define a distance function on \mathcal{V} as

$$d(V_1, V_2) \triangleq \sup_{(x,t) \in G} \{|V_1(x, t) - V_2(x, t)|\} \quad \text{for } V_1, V_2 \in \mathcal{V}.$$

THEOREM 7. *Let T be continuous in $\mathcal{V}^0 \subset \mathcal{V}$. If $\{V^n\}$ is such that $T(V^n) = V^{n+1}$ and $V^n(x, t) \downarrow V^*(x, t)$ as in the above mentioned construction, and if $V^* \in \mathcal{V}^0$, then $V^* = V^0$.*

Proof. With the defined metric, Lemma 3 implies that $V^n \rightarrow V^*$ and $T(V^n) \rightarrow V^*$. Since T is continuous, $T(V^*) = V^*$, so that by Theorem 4, $V^* = V^0$.

Above, we give conditions under which the iteration will converge to the optimal $V^0(x, t)$. Perhaps more important from a practical standpoint is that the successive approximations are monotone decreasing under more general conditions and a gauge can be obtained by finding *lower* bounds on $V^0(x, t)$ by the use of Theorem 1.

5. Quasilinearization and the canonical equations. It is interesting to note that, by Theorem 2, evaluating the successive approximation $V^{n+1} = T(V^n)$ amounts to solving a sequence of *linear* partial differential equations of the first order, since

$$H(x, V_x^{n+1}, t, c(x, V_x^n, t)) + V_t^{n+1} = 0$$

implies

$$(18) \quad \langle V_x^{n+1}, f(x, c(x, V_x^n, t), t) \rangle + V_t^{n+1} = -L(x, c(x, V_x^n, t), t).$$

It is well known [1], [5] that the Hamilton-Jacobi equation $H^0(x, V_x, t) + V_t = 0$ specifies a two point boundary value problem in the canonical equations $\dot{x} = H_p^0$, $\dot{p} = -H_x^0$. Because of their general nonlinearity, numerical integration of the canonical equations presents a formidable problem. Bellman and Kalaba [6] have employed quasilinearization to reduce problems of this type to problems of solving sequences of *linear* two-point boundary value problems.

As one might expect, the characteristic equations associated with (18)

provide a similar formal mechanism of solution. They may be written with $\phi^n(t) \triangleq \phi_{kn}(t)$ as

$$(19) \quad \begin{aligned} \phi^{n+1}(t) &= H_p(\phi^{n+1}, p^{n+1}, t, c(\phi^n, p^n, t)), \\ -\dot{p}^{n+1}(t) &= H_x(\phi^{n+1}, p^{n+1}, t, c(\phi^n, p^n, t)), \end{aligned}$$

or, more explicitly,

$$(20) \quad \begin{aligned} \phi^{n+1}(t) &= f(\phi^{n+1}, c(\phi^n, p^n, t), t), \\ -\dot{p}^{n+1}(t) &= f_x(\phi^{n+1}, c(\phi^n, p^n, t), t)p^{n+1} + L_x(\phi^{n+1}, c(\phi^n, p^n, t), t) \\ &\quad + c_x(\phi^n, p^n, t)[L_u(\phi^{n+1}, c(\phi^n, p^n, t), t) \\ &\quad + f_u(\phi^{n+1}, c(\phi^n, p^n, t), t)p^n], \end{aligned}$$

where f_x , c_x , f_u are Jacobian matrices and L_x , L_u are gradient vectors. It is seen that the second equation is linear in p^{n+1} . The equations may be integrated iteratively by choosing an appropriate admissible control $u(t)$ in place of $c(\phi^1(t), p^1(t), t)$. This amounts to an "approximation in policy space" [2]. Boundary conditions are obtained from the general transversality condition

$$(21) \quad [d\lambda + H^0 dt + \langle p, dx \rangle]_{t_0}^{t_1} = 0,$$

where the differentials are consistent with the side constraints. We do not wish to go further into these matters, but mention them to show connections with other work in the field.

6. Applications.

Example 1. A norm invariant system may be given by

$$\dot{x} = A(x)x + k, \quad \text{where } A + A^T = 0.$$

Let $S = \{(x, t) \mid \|x\| = \theta\}$, $G = \{(x, t) \mid \|x\| \geq \theta\}$, $U = \{k \mid \|k\| \leq 1\}$, $\lambda = 0$, and $L = 1$, where θ is any (small) positive number. Since without control all solutions are on the constant norm surface $\|x\| = \|x_0\|$, it is reasonable to assume that $V^1(x, t) = g(\|x\|)$, where $dg(\alpha)/d\alpha > 0$. It is easy to show that $c(x, p, t) = -p/\|p\|$; consequently, $k^2(x, t) = c(x, V_x^1, t) = -x/\|x\|$. It is then straightforward to show that $V^3(x, t) = V^2(x, t) = \|x\| - \theta$. Therefore, by Theorem 6, $V^0(x, t) = \|x\| - \theta$, and the optimal control law is $k^0(x, t) = -x/\|x\|$.

Example 2. In the first-order linear system

$$\dot{x} = k,$$

with $L = x^2 + k^2$, $G = U = R^1$, $S = \{(x, t) \mid x = 0\}$, the well-known solution is $V^0(x, t) = x^2$, $k^0(x, t) = -x$. If we assume $V^n(x, t) = Kx^2$, then

(18) becomes

$$-KxV_x^{n+1} + x^2 + K^2x^2 = 0;$$

so $V^{n+1}(x, t) = (1 + K^2)x^2/(2K)$, and it is seen that the related sequence converges to $V^0(x, t)$.

Example 3. A simple, but explicitly insolvable minimum time problem is defined by letting $L = 1$, $G = \{(x, t) \mid \|x\| \geq \theta\}$, $S = \{(x, t) \mid \|x\| = \theta\}$, and $U = \{k \mid \|k\| \leq 1\}$ for the system

$$\begin{aligned}\dot{x}_1 &= -x_1 + k_1, \\ \dot{x}_2 &= -2x_2 + k_2.\end{aligned}$$

The small parameter θ should be held to a positive value in order for the theory to apply, but it seems worthwhile to proceed formally with $\theta = 0$ in the interest of simpler expressions. Application of Theorem 1 shows that

$$\log(1 + 2\|x\|)^{1/2} \leq V^0(x, t) \leq \log(1 + \|x\|).$$

Beginning the iteration of Theorem 5 with $V^1(x, t) = \log(1 + \|x\|)$ we obtain $k^2(x, t) = -x/\|x\|$. Evaluation of $V^2 = T(V^1)$ by means of (18) shows that

$$\begin{aligned}V^2(x, t) &= \log(1 + 2\|x\| + x_1^2)^{1/2}, \\ k^3(x, t) &= \frac{-(x_1(1 + \|x\|), x_2)}{[\|x\|^2 + x_1^2\|x\|^2 + 2\|x\|x_1^2]^{1/2}}.\end{aligned}$$

From the latter control law, $V^3(x, t)$ may be evaluated numerically as the time it takes to reach the origin.

Letting $W(x, t) = (1 + 2\|x\|)^{1/2}$, we have

$$W(x, t) \leq V^0(x, t) \leq V^3(x, t) \leq V^2(x, t) \leq V^1(x, t).$$

The $t = 1$ isochrones for these quantities are shown in Fig. 1. As mentioned above, the set of points where $V^3(x, t) = 1$ was determined by numerical integration. Actually, the set of points satisfying $V^0(x, t) = 1$ in the first quadrant can be specified parametrically by $x_1 = (1/\beta) \log[\beta e + \sqrt{1 + \beta^2 e^2}]$, $x_2 = [e\sqrt{1 + \beta^2 e^2} - \sqrt{1 + \beta^2} - x_1]/(2\beta)$, where β is a parameter ranging over the real line. A comparison shows that the approximation $V^3(x, t) = 1$ is accurate to within a few tenths of a percent.

Example 4. Here we consider a class of first-order problems described by $\dot{x} = g(x) + k$, with $L(x, k, t) = \alpha(x) + k^2$, $G = \{(x, t) \mid x \geq 0\}$, $U = R^1$, $S = \{(x, t) \mid x = 0\}$. We assume that $\alpha(x)$ is positive definite.

In this instance it is possible to carry out the complete approximation process in the policy space \mathcal{K}^0 by using the transformation $T^* = T_1 T_2$

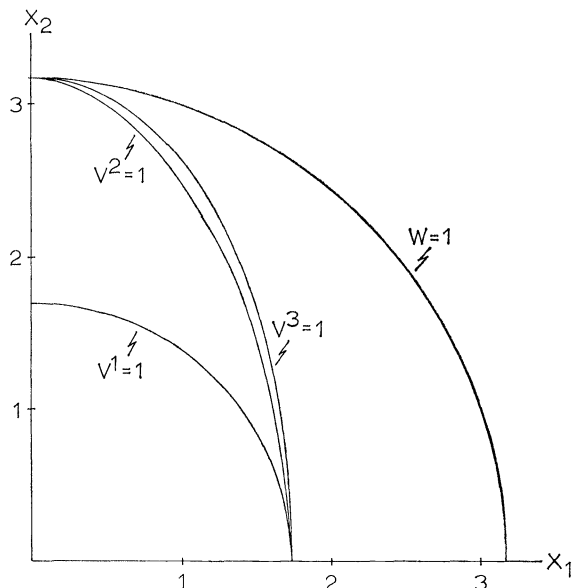


FIG. 1. Isochrones corresponding to $t = 1$ for the successive approximations of Example 3

rather than $T = T_2 T_1$ as has been discussed in the text. We have $c(x, V_x, t) = -\frac{1}{2}V_x$, and (18) becomes

$$(22) \quad V_x^{n+1}(g - \frac{1}{2}V_x^n) + \frac{1}{4}(V_x^n)^2 + \alpha = 0.$$

Further, if $V^n \in \mathfrak{U}^0$, we have by Theorem 1

$$(23) \quad V_x^n g - \frac{1}{4}(V_x^n)^2 + \alpha \leq 0, \quad x \geq 0.$$

Substituting $k^{n+1} = -\frac{1}{2}V_x^n$ into (22) yields

$$k^{n+1} = T^*(k^n) = \frac{1}{2} \frac{(k^n)^2 + \alpha}{k^n + g},$$

or

$$k^{n+1} + g = \frac{1}{2} \left[(k^n + g) + \frac{\alpha + g^2}{(k^n + g)} \right].$$

Now if k^n is admissible, $k^n + g < 0$ for $x > 0$ and k^{n+1} is continuously differentiable for $x > 0$. Furthermore, the origin is asymptotically stable for $\dot{x} = g(x) + k^{n+1}$, since from (23)

$$\dot{V}^n = gV_x^n + k^{n+1}V_x^n = gV_x^n - \frac{1}{2}(V_x^n)^2 \leq -\alpha - \frac{1}{4}(V_x^n)^2,$$

and thus V^n is a Liapunov function for the system. Therefore, k^{n+1} is

admissible. It is easy to show that

$$k^{n+1} + g \rightarrow -\sqrt{\alpha + g^2}, \quad \text{or} \quad k^n \rightarrow -g - \sqrt{\alpha + g^2} = k^0.$$

REFERENCES

- [1] R. E. KALMAN, *The theory of optimal control and the calculus of variations*, Mathematical Optimization Techniques, R. Bellman, ed., University of California Press, Berkeley, 1963, Chap. 16.
- [2] R. E. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.
- [3] Z. V. REKASIUS, *Suboptimal design of intentionally nonlinear controllers*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 380-386.
- [4] R. L. HAUSSLER, *On the sub-optimal design of nonlinear control systems*, Doctoral thesis, Purdue University, Lafayette, 1963.
- [5] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Interscience, New York, 1962.
- [6] R. E. BELLMAN AND R. E. KALABA, *Quasilinearization and Nonlinear Boundary-Value Problems*, American Elsevier, New York, 1965.

CONTROLLABILITY AND OBSERVABILITY IN TIME-VARIABLE LINEAR SYSTEMS*

L. M. SILVERMAN† AND H. E. MEADOWS‡

Introduction. It is well known [1], [2], [3] that the controllability and observability properties of a time-variable linear system may be fully determined from the system solution. However, the solution to such a system is generally not available in closed form. In this paper, we examine the extent to which the various types of controllability and observability may be characterized in terms of the known system coefficient matrices. Of prime importance in this development are the system controllability and observability matrices to be defined below. It will be shown that these matrices, which do not require knowledge of the system solution for their construction, provide significant structural information including a necessary and sufficient condition for total [3] (differential [2]) controllability and observability. This condition is established by relating the controllability and observability matrices to a new test for linear independence of vector functions. This test is a generalization of the familiar Wronskian determinant test for scalar functions. A new degree of controllability and observability is introduced here, and its relation to previous definitions is discussed. It is shown that the state of a system possessing these properties can be controlled and observed “instantaneously.”

Preliminaries. The class of systems to be considered in this paper are those describable by a finite set of first order differential equations of the form¹

$$(1a) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

$$(1b) \quad y(t) = C(t)x(t),$$

* Received by the editors August 31, 1966.

This research was partially supported by the National Science Foundation under Grant NSF GP-2789, and by the Office of Naval Research under Contract NONR 4259(04).

† Department of Electrical Engineering, University of California, Berkeley, California. Formerly with the Department of Electrical Engineering, Columbia University, New York, New York.

‡ Department of Electrical Engineering, Columbia University, New York, New York.

¹ Lower case symbols will be used to denote vector functions, while upper case symbols will be reserved for all other matrices. The operations of transposition and inversion on a matrix A will be denoted A' and A^{-1} respectively and the notations $dA(t)/dt = \dot{A}(t) = A^{(1)}(t)$ and $d^k A(t)/dt^k = A^{(k)}(t)$ will be used for differentiation.

where $x(t)$, an n -vector, is the state of the system of time t ; $u(t)$, an r -vector, is the input; and $y(t)$, an m -vector, is the output. The matrices $A(t)$, $B(t)$ and $C(t)$ are of order compatible with the vectors $x(t)$, $u(t)$ and $y(t)$. To avoid unnecessary complication, it will be assumed that $A(t)$, $B(t)$ and $C(t)$, together with their first $n - 2$, $n - 1$, and $n - 1$ derivatives respectively, are continuous functions.

As is well known [4], the output of such a system is given by

$$(2) \quad y(t) = C(t)X(t) \left[X^{-1}(t_0) x(t_0) + \int_{t_0}^t X^{-1}(\tau)B(\tau)u(\tau) d\tau \right],$$

where $x(t_0)$ is the state of the system at some arbitrary time t_0 , and $X(t)$ is a fundamental matrix [4] for the homogeneous part of (1a), that is,

$$(3) \quad \dot{X}(t) = A(t)X(t)$$

and $X(t)$ is nonsingular for all t .

A brief summary of the important types of controllability and observability for time-variable linear systems will now be given, together with the necessary and sufficient conditions for each in terms of $X(t)$. Various equivalent definitions and criteria, as well as discussion of the role they play in optimal control and filtering theory, may be found in [1], [2] and [3].

DEFINITION 1. (a) System (1) is *completely controllable* on an interval² (t_0, t_1) if, for any initial state x_0 at t_0 , and any desired final state x_1 at t_1 , there exists an input $u(t)$ defined on (t_0, t_1) such that $x(t_1) = x_1$.

(b) System (1) is *totally controllable* [3] (*differentially controllable* [2]) on an interval (t_0, t_1) if it is completely controllable on every subinterval of (t_0, t_1) .

DEFINITION 2. (a) System (1) is *completely observable* on an interval (t_0, t_1) if any initial state x_0 at t_0 can be determined from knowledge of the system output $y(t)$ and input $u(t)$ over (t_0, t_1) .

(b) System (1) is *totally observable* [3] (*differentially observable* [2]) on an interval (t_0, t_1) if it is completely observable on every subinterval of (t_0, t_1) .

If we let $\Theta(t) = X^{-1}(t)B(t)$ and $\Psi(t) = C(t)X(t)$, then the well-known [2], [3] criteria for the above types of controllability and observability are given in the following theorems.

THEOREM 1. (a) System (1) is *completely controllable* on the interval (t_0, t_1) if and only if the rows of $\Theta(t)$ are linearly independent on (t_0, t_1) .

(b) System (1) is *totally controllable* on the interval (t_0, t_1) if and only if the rows of $\Theta(t)$ are linearly independent on every subinterval of (t_0, t_1) .

THEOREM 2. (a) System (1) is *completely observable* on the interval (t_0, t_1) if and only if the columns of $\Psi(t)$ are linearly independent on (t_0, t_1) .

² All intervals considered are open.

(b) *System (1) is totally observable on the interval (t_0, t_1) if and only if the columns of $\Psi(t)$ are linearly independent on every subinterval of (t_0, t_1) .*

The controllability and observability matrices. The necessary and sufficient conditions for controllability and observability summarized in the previous section depend explicitly on $X(t)$, which for time-variable differential equations is rarely available save in numerically tabulated form. In order to circumvent the problem of solving time-variable differential equations, we propose to determine the extent to which the controllability and observability properties can be characterized in terms of the matrices $A(t)$, $B(t)$ and $C(t)$. To this end, we define the controllability matrix of system (1):

$$(4) \quad Q_c(t) = [P_0(t) : P_1(t) : \cdots : P_{n-1}(t)],$$

where

$$(5) \quad P_{k+1}(t) = -A(t)P_k(t) + \dot{P}_k(t), \quad P_0(t) = B(t).$$

The observability matrix is defined similarly:

$$(6) \quad Q_0(t) = [S_0(t) : S_1(t) : \cdots : S_{n-1}(t)],$$

where

$$(7) \quad S_{k+1}(t) = A'(t)S_k(t) + \dot{S}_k(t), \quad S_0(t) = C'(t).$$

To indicate the role played by the controllability matrix as a test for controllability it is first noted that

$$(8) \quad \frac{d^k}{dt^k} [X^{-1}(t)B(t)] = X^{-1}(t)P_k(t),$$

as can be shown by a simple induction argument [5]. Thus,

$$(9) \quad [\Theta(t) : \Theta^{(1)}(t) : \cdots : \Theta^{(n-1)}(t)] = X^{-1}(t)Q_c(t).$$

For a single-input system, the matrix on the left-hand side of (9) is recognized to be the Wronskian matrix [6] of the rows of $\Theta(t)$. Since $X(t)$ is nonsingular for all t , the rank of $Q_c(t)$ is equal to that of the Wronskian matrix for all t . For single-input systems, therefore, it is clear that the controllability matrix yields as much information about the degree of controllability of the system as does the Wronskian matrix of the rows of $\Theta(t)$. Observe, however, that the former matrix may be constructed *without* knowledge of the system solution.

In order to utilize (9) fully, it is necessary to consider a generalization of the usual Wronskian matrix. In the following section, a Wronskian matrix for vector functions will be defined, and its utility as a test for linear independence of such functions will be established.

The vector Wronskian matrix. Consider the set of r -dimensional row vector functions $f_1(t), f_2(t), \dots, f_n(t)$, the elements of which, together with their first $n - 1$ derivatives, are continuous functions. The Wronskian matrix of such a set of functions is defined as

$$W(f_1(t), f_2(t), \dots, f_n(t)) = [F(t) : F^{(1)}(t) : \dots : F^{(n-1)}(t)],$$

where

$$F(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

When the context is clear, $W(t)$ will denote the Wronskian matrix. Several basic properties of this vector Wronskian matrix will now be developed. The first is a direct generalization of the familiar test for linear independence of scalar functions.

LEMMA 1. *If $W(t)$ has rank n for some $t \in (t_0, t_1)$, then the vector functions $f_1(t), f_2(t), \dots, f_n(t)$ are linearly independent on (t_0, t_1) .*

Proof. Suppose the functions are dependent on (t_0, t_1) ; then there exists a constant nonzero n -vector λ such that for all $t \in (t_0, t_1)$, $\lambda'F(t) = 0$. By differentiating this relationship $n - 1$ times, it is clear that $\lambda'W(t) = 0$ for all $t \in (t_0, t_1)$, which contradicts $W(t)$ having rank n for some $t \in (t_0, t_1)$.

Less obvious is the following test for linear dependence of vector functions.

LEMMA 2. *If the Wronskian matrix of the vector functions $f_1(t), f_2(t), \dots, f_n(t)$ has rank $q < n$ for all $t \in (t_0, t_1)$, and if the Wronskian matrix of any q of the functions also has rank q on the interval, then the n functions are linearly dependent on (t_0, t_1) and may be expressed as a linear combination of the q independent functions.*

Although the statement of this lemma is quite similar to a well-known result for scalar functions [6], it is not possible to generalize the usual proof, which relies on the relation the Wronskian of scalar functions bears to the solutions of differential equations. The following proof is believed to be novel even when specialized to the scalar case.

Proof of Lemma 2. Without loss of generality it may be assumed that the Wronskian matrix of $f_1(t), f_2(t), \dots, f_q(t)$ has rank q for all $t \in (t_0, t_1)$. Partition $W(t)$ as

$$W(t) = \begin{bmatrix} W_1(t) \\ \vdots \\ W_2(t) \end{bmatrix},$$

where $W_1(t)$ is the submatrix formed from the first q rows of $W(t)$. Further

partition $W_1(t)$ as

$$W_1(t) = [W_{11}(t) : W_{12}(t)],$$

where $W_{11}(t)$ is the submatrix formed from the first rq columns of $W_1(t)$, that is,

$$W_{11}(t) = W(f_1(t), f_2(t), \dots, f_q(t)).$$

Also let $W_2(t)$ be partitioned conformably with $W_1(t)$ as

$$W_2(t) = [W_{21}(t) : W_{22}(t)].$$

Since $W(t)$ and $W_{11}(t)$ both have rank q for all $t \in (t_0, t_1)$,

$$(10) \quad W_2(t) = K(t)W_1(t)$$

on the interval, where

$$K(t) = W_{21}(t)W'_{11}(t)(W_{11}(t)W'_{11}(t))^{-1}.$$

All that need be shown to prove the lemma is that $K(t)$ is a constant matrix, since it follows trivially from (10) that

$$\begin{bmatrix} f_{q+1}(t) \\ f_{q+2}(t) \\ \vdots \\ f_n(t) \end{bmatrix} = K(t) \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_q(t) \end{bmatrix}.$$

To establish that $K(t)$ is constant, it is first noted that (10) implies

$$(11) \quad \dot{W}_{21}(t) = \dot{K}(t)W_{11}(t) + K(t)\dot{W}_{11}(t).$$

It is clear, however, that $\dot{W}_{21}(t)$ and $\dot{W}_{11}(t)$ are corresponding submatrices of $\dot{W}_2(t)$ and $\dot{W}_1(t)$, respectively, so that (10) implies

$$\dot{W}_{21}(t) = K(t)\dot{W}_{11}(t).$$

Thus, $K(t)W_{11}(t) = 0$ for all $t \in (t_0, t_1)$. Since $W_{11}(t)$ has q rows and rank q everywhere on the interval, it must be true that $\dot{K}(t) = 0$ for all $t \in (t_0, t_1)$. This completes the proof.

Based on Lemmas 1 and 2, we may now establish a result which will be of prime importance in characterizing total controllability and observability.

LEMMA 3. *A necessary and sufficient condition for the vector functions $f_1(t), f_2(t), \dots, f_n(t)$ to be linearly independent on every subinterval of (t_0, t_1) is that their Wronskian matrix not have rank less than n on any subinterval of (t_0, t_1) (i.e., $W(t)$ has rank n on a set of points everywhere dense in (t_0, t_1)).*

Proof. See the Appendix.

Criteria for controllability and observability. Consider now the Wronskian matrix of the rows of $\Theta(t)$,

$$W(\theta_1(t), \theta_2(t), \dots, \theta_n(t)) = [\Theta(t) : \Theta^{(1)}(t) : \dots : \Theta^{(n-1)}(t)],$$

which from (9) can be written as

$$(12) \quad W(t) = X^{-1}(t)Q_c(t).$$

As noted previously, the rank of $W(t)$ is at all times equal to that of $Q_c(t)$. Immediate application of Lemma 1, therefore, gives the following sufficient condition for complete controllability.

THEOREM 3. *System (1) is completely controllable on (t_0, t_1) if $Q_c(t)$ has rank n for some $t \in (t_0, t_1)$.*

This result was also proved directly in [7] and independently by Stubberud [8] and Chang [9].

In general, the condition of Theorem 3 is *not* necessary for complete controllability, just as the vanishing of the Wronskian determinant does not imply the linear dependence of a set of scalar functions. A simple example of a system which demonstrates this fact was given in [7]. As should be clear from Lemma 3, however, a necessary and sufficient condition for total controllability can be given in terms of the matrix $Q_c(t)$. This condition, which constitutes the major result of this paper, is given below.

THEOREM 4. *System (1) is totally controllable on the interval (t_0, t_1) if and only if $Q_c(t)$ does not have rank less than n on any subinterval of (t_0, t_1) .*

By duality [1], the following criteria for observability also hold.

THEOREM 5. *System (1) is completely observable on the interval (t_0, t_1) if $Q_0(t)$ has rank n for some $t \in (t_0, t_1)$.*

THEOREM 6. *System (1) is totally observable on the interval (t_0, t_1) if and only if $Q_0(t)$ does not have rank less than n on any subinterval of (t_0, t_1) .*

In the special case where the coefficient matrices of system (1) are analytic functions of time, the distinction between complete and total controllability disappears, since if a set of analytic functions are linearly independent on any subinterval of a given interval they are linearly independent on every subinterval of the interval. It also follows from the elementary properties of analytic functions that an analytic matrix has constant rank save possibly at a finite number of points over any finite interval.³ Thus, directly from Theorem 4, we have the following corollary.

COROLLARY 1.⁴ *An analytic system of the form (1) is completely controllable on the interval (t_0, t_1) if and only if $Q_c(t)$ has rank n on the interval.*

Similarly, Theorem 5 implies the next corollary.

³ For convenience, if an analytic matrix has rank q at all but a finite number of points on an interval, it will be said that the matrix "has rank q " on the interval.

⁴ A similar result for this case was also established by Stubberud [8] and Chang [9].

COROLLARY 2. *An analytic system of the form (1) is completely observable on the interval (t_0, t_1) if and only if $Q_0(t)$ has rank n on the interval.*

While complete and total controllability and observability play an important role in optimal control and filtering theory, it has recently been noted [10], [11] that somewhat stronger properties are necessary for certain problems involving system transformations of coordinates. These properties, termed uniform controllability and uniform observability, are defined below.

DEFINITION 3. System (1) is said to be *uniformly controllable* on the interval (t_0, t_1) if the matrix $Q_c(t)$ has rank n for all $t \in (t_0, t_1)$.

DEFINITION 4. System (1) is said to be *uniformly observable* on the interval (t_0, t_1) if the matrix $Q_0(t)$ has rank n for all $t \in (t_0, t_1)$.

It is clear (even for analytic systems) that uniform controllability is a stronger property than total controllability. An interesting interpretation of uniform controllability can be made which shows how this criterion relates to the more familiar types of controllability arising in optimal control problems.

If a system is totally controllable, then by definition the state of the system may be transferred to any desired value in an arbitrarily short interval of time by application of some input. It will now be shown that if a system is uniformly controllable, it is possible to perform the state transition "instantaneously." Furthermore, an explicit input in terms of the controllability matrix will be given which effects the state transition.

Suppose the state of system (1) is zero at time t_0 and $u(t) = \delta(t - t_0)\alpha_0$, where $\delta(t - t_0)$ is the unit impulse "function" applied at t_0 , and α_0 is an r -vector of arbitrary constants. Then, by using the well-known properties of the impulse function [12],

$$x(t) = X(t)X^{-1}(t_0)B(t_0)\alpha_0, \quad t > t_0.$$

Thus it may be said that the state at time t_0 has changed instantaneously from zero to

$$(13) \quad x(t_0) = B(t_0)\alpha_0 = P_0(t_0)\alpha_0.$$

By generalizing the argument of Zadeh and Desoer [10, p. 496] to time-variable systems and utilizing (8) it can be seen that if $u(t) = \delta^{(k)}(t - t_0)\alpha_k$ is applied to system (1) the state will "jump" to the value

$$(14) \quad x(t_0) = P_k(t_0)\alpha_k,$$

where $P_k(t)$ is defined by the recursion formula (5).

Suppose now that x_i is the initial state of the system at t_0 , and x_d is the

desired final state. It is clear that, if the input is of the form

$$(15) \quad u(t) = \sum_{k=0}^{n-1} \delta^{(k)}(t - t_0) \alpha_k,$$

then

$$(16) \quad (x_d - x_i) = \sum_{k=0}^{n-1} P_k(t_0) \alpha_k.$$

If the nr -vector α is defined as

$$\alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_{n-1} \end{bmatrix},$$

then (16) can be rewritten as

$$(17) \quad (x_d - x_i) = Q_c(t_0) \alpha.$$

Equation (17) represents n equations in nr unknowns, and if $Q_c(t_0)$ has rank n , a solution (nonunique, in general, if $r > 1$) for α exists. If we let $G(t_0)$ be the matrix formed from the first n independent columns of $Q_c(t_0)$ and $\bar{\alpha}$ be the vector formed from the corresponding elements of α , then a solution which minimizes the order of the impulse functions required is obtained by setting all the elements of α equal to zero save those in $\bar{\alpha}$, which are given by

$$(18) \quad \bar{\alpha} = G^{-1}(t_0)(x_d - x_i).$$

A dual interpretation for uniform observability may also be made. Let $u(t) = 0$ for convenience, and differentiate the output of system (1) $n - 1$ times. It is clear that

$$(19) \quad Y(t) \triangleq \begin{bmatrix} y(t) \\ y^{(1)}(t) \\ \vdots \\ y^{(n-1)}(t) \end{bmatrix} = Q_0'(t)x(t).$$

If $Q_0(t_0)$ has rank n , then (19) can be solved uniquely for $x(t_0)$, with the solution given explicitly as

$$(20) \quad x(t_0) = [Q_0'(t_0)]^\dagger Y(t_0),$$

where

$$[Q_0'(t_0)]^\dagger = [Q_0(t_0)Q_0'(t_0)]^{-1}Q_0(t_0).$$

That is, if system (1) is uniformly observable, the state of the system at

any time may be determined instantaneously from observations of the output and its derivatives.

It should be noted that one application of uniform observability is to the problem, discussed by Weiss [2], of obtaining a minimal set of differential equations directly relating the output to the system input. It is easily shown that uniform rather than differential observability is the required necessary condition for the existence of such a set of equations. As shown in [10], furthermore, it is not necessary to compute the system solution in order to obtain the equations. They may be constructed directly from $Q_0(t)$.

Appendix. Proof of Lemma 3. Sufficiency follows immediately from Lemma 1.

Let I be any subinterval of (t_0, t_1) on which $W(t)$ has rank less than n for all t . To establish necessity it will be shown by induction on n that there exists a subinterval of I over which the given functions are linearly dependent. This procedure is a generalization of that for the scalar case given by Hurewicz [6].

For $n = 1$, the argument is trivial. Assume then for some $k > 1$ that the statement is valid for $n = k - 1$, and suppose that $W(f_1(t), f_2(t), \dots, f_k(t))$ has rank less than k for all $t \in I$. Then, either $W(f_1(t), f_2(t), \dots, f_{k-1}(t))$ has rank $k - 1$ for some $t \in I$, or it has rank less than $k - 1$ for all $t \in I$. If the former is true, then by continuity there must exist some subinterval $J \subset I$ on which $W(f_1(t), f_2(t), \dots, f_{k-1}(t))$ has rank $k - 1$ everywhere. In this case, Lemma 2 implies that $f_1(t), f_2(t), \dots, f_k(t)$ are dependent on J . If $W(f_1(t), f_2(t), \dots, f_{k-1}(t))$ has rank less than $k - 1$ for all $t \in I$, then by the induction assumption $f_1(t), f_2(t), \dots, f_{k-1}(t)$ are linearly dependent on some subinterval of I . This completes the proof.

REFERENCES

- [1] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.
- [2] L. WEISS, *The concepts of differential controllability and differential observability*, J. Math. Anal. Appl., 10 (1965), pp. 442-449.
- [3] E. KREINDLER AND P. E. SARACHIK, *On the concepts of controllability and observability of linear systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 129-136.
- [4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [5] L. M. SILVERMAN, *Representation and realization of time-variable linear systems*, Tech. Rept. 94, Department of Electrical Engineering, Columbia University, New York, 1966.
- [6] W. HUREWICZ, *Lectures on Ordinary Differential Equations*, M.I.T. Press, Cambridge, 1958.
- [7] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and time-variable unilateral networks*, IEEE Trans. Circuit Theory, CT-12 (1965), pp. 308-313.

- [8] A. R. STUBBERUD, *A controllability criterion for a class of linear systems*, IEEE Trans. Applications and Industry, 68 (1964), pp. 411-413.
- [9] A. CHANG, *An algebraic characterization of controllability*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 112-113.
- [10] L. M. SILVERMAN AND H. E. MEADOWS, *Degrees of controllability in time-variable linear systems*, Proceedings of the National Electronics Conference, 21 (1965), pp. 689-693.
- [11] L. M. SILVERMAN, *Transformation of time-variable systems to canonical (phase-variable) form*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 300-303.
- [12] L. A. ZADEH AND C. A. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1963.

ORTHOGONAL PROJECTION AND DISCRETE OPTIMAL LINEAR SMOOTHING*

JAMES S. MEDITCH†

Abstract. The smoothing filter and smoothing error covariance matrix equations are developed for discrete linear systems using the method of orthogonal projection. Two equivalent formulations are presented and found to agree with those of previous authors who had used other methods. The present results in conjunction with the earlier work of Kalman on prediction and filtering give a complete treatment of the discrete linear estimation problem from the viewpoint of orthogonal projection.

1. Introduction. In 1960 Kalman [1] utilized the method of orthogonal projection [2], [3] to develop the fundamental equations of optimal linear filtering and prediction for discrete linear systems. The following year, he and Bucy [4] applied the same principle to derive the analogous equations for continuous linear systems. Subsequently, other authors have approached the estimation problem, i.e., prediction, filtering, and smoothing, from a number of viewpoints. Among the methods used are least squares [5], maximum likelihood [6], [7], linear regression [8], dynamic programming [9], stochastic approximation [10], and the Bayesian approach [11].

Almost exclusively, attention has been devoted to the prediction and filtering problems, whereas the smoothing problem has received comparatively little study. Noteworthy exceptions are the efforts of Bryson and Frazier [12] utilizing the calculus of variations, and Rauch, Tung, and Striebel [7] utilizing maximum likelihood.

In his paper [1], Kalman formulated the entire estimation problem in terms of orthogonal projection, but considered only the prediction and filtering aspects. In this paper, we develop the fundamental equations of optimal linear smoothing for discrete linear systems using orthogonal projection. Our development along with Kalman's [1] then provides a unified treatment of the estimation problem from the viewpoint of orthogonal projection. We remark that these results can readily be extended to include estimation for continuous linear systems by utilizing appropriate limiting processes [13].

2. Optimal linear estimation and orthogonal projection. Since our work is based on Kalman's [1] original formulation of the estimation problem in terms of orthogonal projection, we begin with a review of his work. We

* Received by the editors May 17, 1966.

† Department of Electrical Engineering, Northwestern University, Evanston, Illinois 60201. This work was supported in part by the U. S. Army, Navy, and Air Force in a Joint Service Electronics Program under Office of Naval Research Contract N 00014-66-C0020-A03.

consider the system

$$(1) \quad x(k+1) = \Phi(k+1, k)x(k) + u(k),$$

$$(2) \quad z(k+1) = H(k+1)x(k+1) + v(k+1),$$

where x is an n -vector, the state; z is an m -vector, the output (measurement); Φ is a real $n \times n$ matrix, the state transition matrix; H is a real $m \times n$ matrix; and $k = 0, 1, \dots$ is the discrete time index. In addition, u and v are independent, zero mean, Gaussian white sequences for which

$$E[u(j)u'(k)] = Q(k)\delta_{jk},$$

$$E[v(j)v'(k)] = R(k)\delta_{jk},$$

where E denotes the expected value, the prime denotes the matrix transpose, and δ_{jk} is the Kronecker delta. Here, Q is a real $n \times n$ positive semidefinite matrix and R is a real $m \times m$ positive definite matrix. The initial state $x(0)$ is assumed to be a zero mean Gaussian random n -vector which is independent of u and v for all k , and for which

$$E[x(0)x'(0)] = P(0),$$

where $P(0)$ is a real $n \times n$ positive semidefinite matrix.

An estimate of $x(k)$ based on measurements up to and including the one at time j will be denoted by $\hat{x}(k|j)$. If $k = j$, the notation $\hat{x}(k) \triangleq \hat{x}(k|k)$ will be used. The three basic problems in estimation are: (1) prediction, $k > j$; (2) filtering, $k = j$; and (3) smoothing, $k < j$.

The estimation error is defined by the relation

$$\tilde{x}(k|j) = x(k) - \hat{x}(k|j).$$

An estimate of the form

$$\hat{x}(k|j) = \sum_{i=1}^j A(i)z(i),$$

where the $A(i)$ are real $n \times m$ matrices, which minimizes the mean square estimation error

$$E[\tilde{x}'(k|j)\tilde{x}(k|j)] = E\left[\sum_{i=1}^n \tilde{x}_i^2(k|j)\right],$$

will be called an *optimal estimate* of $x(k)$.

Now let $Y(j)$ denote the vector space which is generated by considering the set of all linear combinations of the form

$$\sum_{i=1}^j B(i)z(i),$$

where the $B(i)$ are arbitrary real $n \times m$ matrices.

The orthogonal projection of $x(k)$ on $Y(j)$, denoted by $\bar{x}(k|j)$, has the following three properties [1], [15]:

- (i) $\bar{x}(k|j) \in Y(j)$;
- (ii) $x(k) - \bar{x}(k|j)$ is orthogonal¹ to $Y(j)$;
- (iii) if $x(k) - \xi$ is orthogonal to $Y(j)$ and $\xi \in Y(j)$, then $\xi = \bar{x}(k|j)$.

The next two theorems then follow immediately.

THEOREM 1 (Kalman). *The optimal estimate of $x(k)$ is the orthogonal projection of $x(k)$ on $Y(j)$.*

THEOREM 2 (Kalman). *If $\hat{x}(k)$ is given, then*

$$(3) \quad \hat{x}(k+1|k) = \Phi(k+1, k)\hat{x}(k), \quad k = 0, 1, \dots$$

Theorem 2 gives the solution of the prediction problem. It is then easy to show that the prediction error covariance matrix $M(k+1)$ is given by the expression

$$(4) \quad M(k+1) = \Phi(k+1, k)P(k)\Phi'(k+1, k) + Q(k), \quad k = 0, 1, \dots,$$

where

$$M(k+1) = E[\bar{x}(k+1|k)\bar{x}'(k+1|k)]$$

and

$$P(k) = E[\bar{x}(k)\bar{x}'(k)],$$

where the latter is the filtering error covariance matrix.

We now let $\hat{z}(k|j) \triangleq H(k)\hat{x}(k|j)$ and $\bar{z}(k|j) \triangleq z(k) - \hat{z}(k|j)$. We note that

$$(5) \quad \bar{z}(k+1|k) = z(k+1) - H(k+1)\hat{x}(k+1|k)$$

$$(6) \quad = z(k+1) - H(k+1)\Phi(k+1, k)\hat{x}(k).$$

Letting $Z(k+1)$ denote the vector space which is generated by considering the set of all n -vectors of the form

$$K(k+1)\bar{z}(k+1|k),$$

where $K(k+1)$ is any arbitrary real $n \times m$ matrix, we can easily show [1], [14] that $Y(k)$ and $Z(k+1)$ are mutually orthogonal vector spaces, i.e., all vectors in $Y(k)$ are orthogonal to all vectors in $Z(k+1)$. In addition, using o.p. $[x, Y]$ to denote the orthogonal projection of x on Y , it can be

¹Two random vectors a and b are orthogonal if their inner product $\langle a, b \rangle = E(b'a)$ vanishes. A vector is orthogonal to a vector space if it is orthogonal to every vector in the space. Since we have only zero mean Gaussian random vectors here, we have the result that orthogonal zero mean Gaussian random vectors are uncorrelated, and, therefore, independent [1].

shown [1], [14] that

$$\begin{aligned} Y(k+1) &= Y(k) + Z(k+1), \\ \text{o.p. } [x(k+1), Y(k+1)] &= \text{o.p. } [x(k+1), Y(k)] \\ &\quad + \text{o.p. } [x(k+1), Z(k+1)]. \end{aligned}$$

This leads to the following theorem.

THEOREM 3 (Kalman). *The optimal filtered estimate of $x(k+1)$ is given by the relation*

$$\begin{aligned} (7) \quad \hat{x}(k+1) &= \hat{x}(k+1|k) + K^0(k+1)\bar{z}(k+1|k) \\ &= \Phi(k+1, k)\hat{x}(k) \\ (8) \quad &+ K^0(k+1)[z(k+1) - H(k+1)\Phi(k+1, k)\hat{x}(k)]. \end{aligned}$$

The optimal filter gain $K^0(k+1)$ is obtained from the relations [14]:

$$(9) \quad K^0(k+1) = M(k+1)H'(k+1)[H(k+1)M(k+1)H'(k+1) + R(k+1)]^{-1},$$

$$(10) \quad M(k+1) = \Phi(k+1, k)P(k)\Phi'(k+1, k) + Q(k),$$

$$\begin{aligned} (11) \quad P(k+1) &= [I - K^0(k+1)H(k+1)]M(k+1)[I - K^0(k+1) \\ &\quad \cdot H(k+1)]' + K^0(k+1)R(k+1)K^{0'}(k+1) \\ &= [I - K^0(k+1)H(k+1)]M(k+1) \end{aligned}$$

for $k = 0, 1, \dots$, where $[\]^{-1}$ denotes the matrix inverse, I is the $n \times n$ identity matrix, and M and P are as defined above. The initial values in (8) and (10) are $\hat{x}(0) = 0$ and $P(0)$, respectively.

3. Single and double stage smoothing. We are now in a position to solve the smoothing problem. Our development is by induction. Hence, we begin by first obtaining $\hat{x}(k|k+1)$ and $\hat{x}(k|k+2)$, which we term the single and double stage smoothed estimates, respectively.

Single stage. From Theorem 1 and the nature of the vector spaces $Y(k+1)$, $Y(k)$, and $Z(k+1)$, we have immediately that

$$\begin{aligned} (12) \quad \hat{x}(k|k+1) &= \text{o.p. } [x(k), Y(k+1)] \\ &= \text{o.p. } [x(k), Y(k)] + \text{o.p. } [x(k), Z(k+1)] \\ &= \hat{x}(k) + K(k+1)\bar{z}(k+1|k), \end{aligned}$$

where $K(k+1)$ is to be determined.

Since $K(k+1)\bar{z}(k+1|k) = \text{o.p. } [x(k), Z(k+1)]$, it follows from the second property of orthogonal projection that $x(k) - K(k+1)\bar{z}(k+1|k)$

is orthogonal to $Z(k+1)$. Furthermore, since $\tilde{z}(k+1|k) \in Z(k+1)$ and orthogonal zero mean Gaussian random vectors are uncorrelated, we have that

$$(13) \quad E[x(k)\tilde{z}'(k+1|k) - K(k+1)\tilde{z}(k+1|k)\tilde{z}'(k+1|k)] = 0.$$

Substituting $x(k) = \hat{x}(k) + \tilde{x}(k)$ into (13), and noting that $\hat{x}(k) \in Y(k)$, $\tilde{z}(k+1|k) \in Z(k+1)$, and that $Y(k)$ and $Z(k+1)$ are mutually orthogonal, we obtain

$$(14) \quad E[\tilde{x}(k)\tilde{z}'(k+1|k) - K(k+1)\tilde{z}(k+1|k)\tilde{z}'(k+1|k)] = 0.$$

From the definitions of $\tilde{x}(k+1|k)$ and $\tilde{z}(k+1|k)$, and (1), (2), and (3), we have

$$\begin{aligned} \tilde{x}(k+1|k) &= x(k+1) - \hat{x}(k+1|k) \\ (15) \quad &= \Phi(k+1, k)x(k) + u(k) - \Phi(k+1, k)\hat{x}(k) \\ &= \Phi(k+1, k)\tilde{x}(k) + u(k), \end{aligned}$$

and

$$\begin{aligned} \tilde{z}(k+1|k) &= z(k+1) - H(k+1)\hat{x}(k+1|k) \\ &= H(k+1)x(k+1) + v(k+1) - H(k+1)\hat{x}(k+1|k) \\ &= H(k+1)\tilde{x}(k+1|k) + v(k+1). \end{aligned}$$

Since u and v are independent zero mean Gaussian white sequences, it can be shown after some manipulation that

$$E[\tilde{x}(k)v'(k+1)] = 0$$

and

$$E[\tilde{x}(k+1|k)v'(k+1)] = 0$$

for all k . It then follows that

$$(16) \quad E[\tilde{x}(k)\tilde{z}'(k+1|k)] = E[\tilde{x}(k)\tilde{x}'(k+1|k)]H'(k+1)$$

and

$$\begin{aligned} (17) \quad E[\tilde{z}(k+1|k)\tilde{z}'(k+1|k)] \\ = H(k+1)M(k+1)H'(k+1) + R(k+1), \end{aligned}$$

where all of the terms have been defined previously.

Utilizing (15), we have that

$$\begin{aligned} (18) \quad E[\tilde{x}(k)\tilde{x}'(k+1|k)] &= E[\tilde{x}(k)\tilde{x}'(k)]\Phi'(k+1, k) + E[\tilde{x}(k)u'(k)] \\ &= P(k)\Phi'(k+1, k), \end{aligned}$$

where we have made use of the readily verified fact that $E[\hat{x}(k)u'(k)] = 0$ for all k . Hence, (16) becomes

$$(19) \quad E[\hat{x}(k)\hat{z}'(k+1|k)] = P(k)\Phi'(k+1, k)H'(k+1).$$

Substituting (17) and (19) in (14), we obtain

$$\begin{aligned} P(k)\Phi'(k+1, k)H'(k+1) \\ - K(k+1)[H(k+1)M(k+1)H'(k+1) + R(k+1)] = 0. \end{aligned}$$

Solving for $K(k+1)$, we have

$$(20) \quad \begin{aligned} K(k+1) &= P(k)\Phi'(k+1, k)H'(k+1) \\ &\cdot [H(k+1)M(k+1)H'(k+1) + R(k+1)]^{-1}. \end{aligned}$$

But from (9), we note that

$$(21) \quad \begin{aligned} H'(k+1)[H(k+1)M(k+1)H'(k+1) + R(k+1)]^{-1} \\ = M^{-1}(k+1)K^0(k+1) \end{aligned}$$

under the assumption that $M^{-1}(k+1)$ exists. We remark that the existence of $M^{-1}(k+1)$ is assured if $Q(k)$ is positive definite.

Substituting (21) into (20) and the result into (12), we obtain

$$(22) \quad \begin{aligned} \hat{x}(k|k+1) \\ = \hat{x}(k) + P(k)\Phi'(k+1, k)M^{-1}(k+1)K^0(k+1)\hat{z}(k+1|k). \end{aligned}$$

But, from (7), the optimal filter equation, we see that

$$K^0(k+1)\hat{z}(k+1|k) = \hat{x}(k+1) - \hat{x}(k+1|k).$$

Hence, (22) becomes

$$(23) \quad \hat{x}(k|k+1) = \hat{x}(k) + J(k)[\hat{x}(k+1) - \hat{x}(k+1|k)],$$

where

$$(24) \quad J(k) \triangleq P(k)\Phi'(k+1, k)M^{-1}(k+1).$$

The $n \times n$ matrix $J(k)$ is termed the smoothing filter gain matrix.

We remark that if $Q(k)$ and $M(k+1) \equiv 0$, i.e., $M^{-1}(k+1)$ is singular, then $\hat{x}(k+1|k) = \Phi(k+1, k)\hat{x}(k) \equiv x(k+1)$, and there is no need for smoothing since a zero covariance estimate of $x(k)$ can be obtained by pre-multiplying $\hat{x}(k+1|k)$ by $\Phi(k, k+1)$. Hence, we shall assume $M(k+1) \neq 0$ in the sequel. Similarly, if $P(k) = 0$, we note that $J(k) = 0$ and $\hat{x}(k|k+1) = \hat{x}(k)$.

It follows from (23) and (24) that

$$(25) \quad \begin{aligned} \hat{x}(k+1|k+2) \\ = \hat{x}(k+1) + J(k+1)[\hat{x}(k+2) - \hat{x}(k+2|k+1)], \end{aligned}$$

where

$$(26) \quad J(k+1) = P(k+1)\Phi'(k+2, k+1)M^{-1}(k+2).$$

Double stage. Again, from Theorem 1 and the nature of the vector spaces $Y(k+2)$, $Y(k+1)$, and $Z(k+2)$, it follows that

$$\begin{aligned} \hat{x}(k|k+2) &= \text{o.p. } [x(k), Y(k+2)] \\ &= \text{o.p. } [x(k), Y(k+1)] + \text{o.p. } [x(k), Z(k+2)]. \end{aligned}$$

But from the single stage smoothing solution,

$$\text{o.p. } [x(k), Y(k+1)] = \hat{x}(k) + J(k)[\hat{x}(k+1) - \hat{x}(k+1|k)].$$

Also,

$$\text{o.p. } [x(k), Z(k+2)] = K(k+2)\tilde{z}(k+2|k+1).$$

Hence,

$$(27) \quad \begin{aligned} \hat{x}(k|k+2) &= \hat{x}(k) + J(k)[\hat{x}(k+1) - \hat{x}(k+1|k)] \\ &\quad + K(k+2)\tilde{z}(k+2|k+1), \end{aligned}$$

where $K(k+2)$ is to be determined.

As in the single stage case above, $x(k) - K(k+2)\tilde{z}(k+2|k+1)$ is orthogonal to $Z(k+2)$, and since $\tilde{z}(k+2|k+1) \in Z(k+2)$, it immediately follows that

$$E[x(k)\tilde{z}'(k+2|k+1) - K(k+2)\tilde{z}(k+2|k+1)\tilde{z}'(k+2|k+1)] = 0.$$

Substituting $x(k) = \hat{x}(k) + \tilde{x}(k)$ into this expression and noting that $\hat{x}(k) \in Y(k) \subset Y(k+1)$, $\tilde{z}(k+2|k+1) \in Z(k+2)$ and that $Y(k+1)$ and $Z(k+2)$ are mutually orthogonal, we can write

$$(28) \quad \begin{aligned} E[\tilde{x}(k)\tilde{z}'(k+2|k+1) \\ - K(k+1)\tilde{z}(k+2|k+1)\tilde{z}'(k+2|k+1)] = 0. \end{aligned}$$

Utilizing the same procedure as that which led to (17), we obtain

$$(29) \quad \begin{aligned} E[\tilde{z}(k+2|k+1)\tilde{z}'(k+2|k+1)] \\ = H(k+2)M(k+2)H'(k+2) + R(k+2), \end{aligned}$$

which gives us the second term in (28).

We now evaluate the first term. First, from the fact that

$$\tilde{z}(k+2|k+1) = H(k+2)\tilde{x}(k+2|k+1) + v(k+2),$$

it follows that

$$(30) \quad E[\tilde{x}(k)\tilde{z}'(k+2|k+1)] = E[\tilde{x}(k)\tilde{x}'(k+2|k+1)]H'(k+2)$$

since $v(k+2)$ is an independent zero mean Gaussian white sequence and it is easily shown that $E[\tilde{x}(k)v'(k+2)] = 0$ for all k .

Utilizing the definitions of $\tilde{x}(k+2|k+1)$ and $\tilde{x}(k+1)$, and (8), we obtain

$$(31) \quad \tilde{x}(k+2|k+1) = \Phi(k+2, k+1)\tilde{x}(k+1) + u(k+1)$$

and

$$(32) \quad \begin{aligned} \tilde{x}(k+1) &= [I - K^0(k+1)H(k+1)]\tilde{x}(k+1|k) \\ &\quad - K^0(k+1)v(k+1). \end{aligned}$$

Substituting (32) into (31), we have

$$(33) \quad \begin{aligned} &\tilde{x}(k+2|k+1) \\ &= \Phi(k+2, k+1)[I - K^0(k+1)H(k+1)]\tilde{x}(k+1|k) \\ &\quad - \Phi(k+2, k+1)K^0(k+1)v(k+1) + u(k+1). \end{aligned}$$

Since $v(k+1)$ and $u(k+1)$ are independent zero mean Gaussian white sequences, it can be shown by direct evaluation that

$$E[\tilde{x}(k)v'(k+1)] = 0$$

and

$$E[\tilde{x}(k)u'(k+1)] = 0$$

for all k . Therefore,

$$(34) \quad \begin{aligned} &E[\tilde{x}(k)\tilde{x}'(k+2|k+1)] \\ &= E[\tilde{x}(k)\tilde{x}'(k+1|k)][I - K^0(k+1)H(k+1)]'\Phi'(k+2, k+1) \\ &= P(k)\Phi'(k+1, k)[I - K^0(k+1)H(k+1)]'\Phi'(k+2, k+1). \end{aligned}$$

Substituting (34) into (30) and the result along with (29) into (28), and solving for $K(k+2)$ gives us

$$(35) \quad \begin{aligned} &K(k+2) \\ &= P(k)\Phi'(k+1, k)[I - K^0(k+1)H(k+1)]'\Phi'(k+2, k+1) \\ &\quad \cdot H'(k+2)[H(k+2)M(k+2)H'(k+2) + R(k+2)]^{-1}. \end{aligned}$$

From (24) and (21), the latter with $k + 1$ replaced by $k + 2$, we have

$$P(k)\Phi'(k + 1, k) = J(k)M(k + 1)$$

and

$$\begin{aligned} H'(k + 2)[H(k + 2)M(k + 2)H'(k + 2) + R(k + 2)]^{-1} \\ = M^{-1}(k + 2)K^0(k + 2), \end{aligned}$$

respectively. Hence,

$$\begin{aligned} K(k + 2) = J(k)M(k + 1)[I - K^0(k + 1)H(k + 1)]' \\ \cdot \Phi'(k + 2, k + 1)M^{-1}(k + 2)K^0(k + 2). \end{aligned}$$

But, from (11),

$$M(k + 1)[I - K^0(k + 1)H(k + 1)]' = P(k + 1),$$

and it follows that

$$K(k + 2) = J(k)P(k + 1)\Phi'(k + 2, k + 1)M^{-1}(k + 2)K^0(k + 2),$$

which, in view of (26), is simply

$$(36) \quad K(k + 2) = J(k)J(k + 1)K^0(k + 2).$$

Substituting (36) into (27) and utilizing the fact that

$$K^0(k + 2)z(k + 2|k + 1) = \hat{x}(k + 2) - \hat{x}(k + 2|k + 1),$$

we obtain

$$\begin{aligned} \hat{x}(k|k + 2) &= \hat{x}(k) + J(k)[\hat{x}(k + 1) - \hat{x}(k + 1|k)] \\ &\quad + J(k)J(k + 1)[\hat{x}(k + 2) - \hat{x}(k + 2|k + 1)] \\ &= \hat{x}(k) - J(k)\hat{x}(k + 1|k) + J(k)\{\hat{x}(k + 1) \\ &\quad + J(k + 1)[\hat{x}(k + 2) - \hat{x}(k + 2|k + 1)]\}. \end{aligned}$$

But from (25), the term in braces is simply $\hat{x}(k + 1|k + 2)$, and it follows immediately that

$$(37) \quad \hat{x}(k|k + 2) = \hat{x}(k) + J(k)[\hat{x}(k + 1|k + 2) - \hat{x}(k + 1|k)].$$

4. General case. We now proceed by induction. For $N \geq k + 3$, where N is an integer, we write

$$(38) \quad \hat{x}(k|N - 1) = \hat{x}(k) + J(k)[\hat{x}(k + 1|N - 1) - \hat{x}(k + 1|k)],$$

and we seek the expression for $\hat{x}(k|N)$.

From Theorem 1, the properties of orthogonal projection, and (38), we

see that

$$\begin{aligned}
 \hat{x}(k | N) &= \text{o.p. } [x(k), Y(N)] \\
 &= \text{o.p. } [x(k), Y(N-1)] + \text{o.p. } [x(k), Z(N)] \\
 (39) \quad &= \hat{x}(k) + J(k)[\hat{x}(k+1 | N-1) - \hat{x}(k+1 | k)] \\
 &\quad + K(N)\tilde{z}(N | N-1),
 \end{aligned}$$

where $K(N)$ remains to be determined.

As in the single and double stage cases, $x(k) - K(N)\tilde{z}(N | N-1)$ is orthogonal to $Z(N)$ and it follows that

$$(40) \quad E[x(k)\tilde{z}'(N | N-1) - K(N)\tilde{z}(N | N-1)\tilde{z}'(N | N-1)] = 0.$$

As above, since $x(k) = \hat{x}(k) + \tilde{x}(k)$, $\hat{x}(k) \in Y(k) \subset Y(N-1)$, $\tilde{z}(N | N-1) \in Z(N)$, and $Y(N-1)$ and $Z(N)$ are mutually orthogonal, we have

$$E[\hat{x}(k)\tilde{z}'(N | N-1)] = 0,$$

and (40) becomes

$$(41) \quad E[\tilde{x}(k)\tilde{z}'(N | N-1) - K(N)\tilde{z}(N | N-1)\tilde{z}'(N | N-1)] = 0.$$

Utilizing identically the same procedure as the one which led to (36), we obtain

$$(42) \quad K(N) = J(k)J(k+1) \cdots J(N-1)K^0(N).$$

Substituting (42) into (39) and rearranging terms gives us

$$\begin{aligned}
 \hat{x}(k | N) &= \hat{x}(k) - J(k)\hat{x}(k+1 | k) + J(k)[\hat{x}(k+1 | N-1) \\
 (43) \quad &\quad + J(k+1)J(k+2) \cdots J(N-1)K^0(N)\tilde{z}(N | N-1)].
 \end{aligned}$$

The term in brackets can be simplified by noting that

$$\begin{aligned}
 \hat{x}(k+1 | N) &= \text{o.p. } [x(k+1), Y(N)] \\
 (44) \quad &= \text{o.p. } [x(k+1), Y(N-1)] + \text{o.p. } [x(k+1), Z(N)] \\
 &= \hat{x}(k+1 | N-1) + \bar{K}(N)\tilde{z}(N | N-1),
 \end{aligned}$$

where $\bar{K}(N)$ does not necessarily equal $K(N)$ since $\bar{K}(N)\tilde{z}(N | N-1) = \text{o.p. } [x(k+1), Z(N)]$ while $K(N)\tilde{z}(N | N-1) = \text{o.p. } [x(k), Z(N)]$.

Now since $x(k+1) - \bar{K}(N)\tilde{z}(N | N-1)$ is orthogonal to $Z(N)$, we utilize the same argument as that which gave us (41) to obtain

$$E[\hat{x}(k+1)\tilde{z}'(N | N-1) - \bar{K}(N)\tilde{z}(N | N-1)\tilde{z}'(N | N-1)] = 0,$$

from which it can be shown that

$$(45) \quad \bar{K}(N) = J(k+1)J(k+2) \cdots J(N-1)K^0(N)$$

by the same procedure as that used to obtain (36) and subsequently (42).

Substituting (45) into (44), we see that

$$\begin{aligned} \hat{x}(k+1|N) &= \hat{x}(k+1|N-1) \\ &\quad + J(k+1)J(k+2) \cdots J(N-1)K^0(N)\bar{z}(N|N-1), \end{aligned}$$

which is precisely the term in brackets in (43). Hence,

$$(46) \quad \hat{x}(k|N) = \hat{x}(k) + J(k)[\hat{x}(k+1|N) - \hat{x}(k+1|k)],$$

where we recall that

$$(47) \quad J(k) = P(k)\Phi'(k+1, k)M^{-1}(k+1).$$

Equations (46) and (47) specify the optimal linear smoothing filter, and they are in agreement with those obtained by Rauch, Tung, and Striebel [7] who used maximum likelihood.

We note that (46) is a system of n first-order linear, "backward" recursive equations which requires the solution of the prediction and filtering problems, viz., $\hat{x}(k+1|k)$ and $\hat{x}(k)$, respectively, as input. The computation is initiated at $k = N-1$ using $\hat{x}(N|N) \triangleq \hat{x}(N)$, $\hat{x}(N-1)$, and $\hat{x}(N|N-1)$, i.e.,

$$\hat{x}(N-1|N) = \hat{x}(N-1) + J(N-1)[\hat{x}(N) - \hat{x}(N|N-1)].$$

We note also that the smoothing filter gain matrix J depends on $P(k)$, the filtering error covariance matrix, and $M(k+1)$, the prediction error covariance matrix. Hence, the smoothing procedure in (46) and (47) requires storage of $\hat{x}(k)$, $\hat{x}(k+1|k)$, $P(k)$ and $M(k+1)$ for $k = 0, 1, \dots, N$. A block diagram for the smoothing filter is given in Fig. 1.

We now develop a second smoothing procedure which differs slightly from the one presented above. We observe from the second line in (39) that

$$(48) \quad \hat{x}(k|N) = \hat{x}(k|N-1) + K(N)\bar{z}(N|N-1).$$

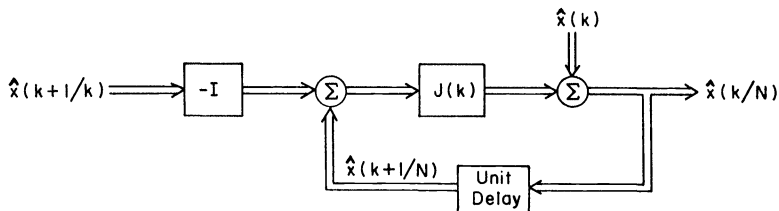


FIG. 1. Block diagram of smoothing solution given by (46)

But from (42),

$$K(N) = J(k)J(k+1) \cdots J(N-1)K^0(N),$$

and from (7) for $k = N-1$,

$$K^0(N)\tilde{z}(N|N-1) = \hat{x}(N) - \hat{x}(N|N-1).$$

Hence, (48) becomes

$$\hat{x}(k|N) = \hat{x}(k|N-1) + \left[\prod_{i=k}^{N-1} J(i) \right] [\hat{x}(N) - \hat{x}(N|N-1)],$$

or

$$(49) \quad \hat{x}(k|N) = \hat{x}(k|N-1) + K(N-1, k)[\hat{x}(N) - \hat{x}(N|N-1)],$$

where

$$(50) \quad K(N-1, k) \triangleq \prod_{i=k}^{N-1} J(i)$$

and

$$(51) \quad J(i) = P(i)\Phi'(i+1, i)M^{-1}(i+1).$$

The distinction to be made between these two formulations is obvious, viz., (46) is a recursive relation between $\hat{x}(k|N)$ and $\hat{x}(k+1|N)$, whereas (49) relates $\hat{x}(k|N)$ and $\hat{x}(k|N-1)$. A block diagram for this alternative procedure is given in Fig. 2.

The results of (49)–(51) agree with those of Rauch [16] who used a different approach.

5. Smoothing error covariance. We develop now the expressions for the smoothing error covariance matrices for the above two smoothing formulations.

From the definition of estimation error and (46), it follows that

$$\begin{aligned} \tilde{x}(k|N) &= x(k) - \hat{x}(k|N) \\ &= \tilde{x}(k) - J(k)[\hat{x}(k+1|N) - \hat{x}(k+1|k)], \end{aligned}$$

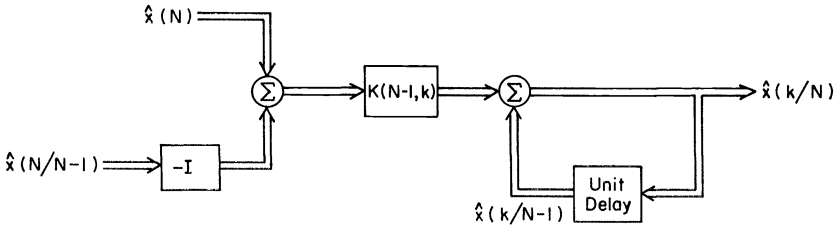


FIG. 2. Block diagram of smoothing solution given by (49)

which we choose to write in the form

$$(52) \quad \tilde{x}(k|N) + J(k)\hat{x}(k+1|N) = \tilde{x}(k) + J(k)\hat{x}(k+1|k).$$

Now we let

$$P(k|N) = E[\tilde{x}(k|N)\tilde{x}'(k|N)],$$

$$P(k) = E[\tilde{x}(k)\tilde{x}'(k)],$$

$$P_{\hat{x}\hat{x}}(k+1|N) = E[\hat{x}(k+1|N)\hat{x}'(k+1|N)],$$

$$P_{\hat{x}\hat{x}}(k+1|k) = E[\hat{x}(k+1|k)\hat{x}'(k+1|k)],$$

where $P(k)$ is as defined earlier. Noting that $\tilde{x}(k|N)$ is orthogonal to $Y(N)$, while $J(k)\hat{x}(k+1|N) \in Y(N)$, and that $\tilde{x}(k)$ is orthogonal to $Y(k)$, while $J(k)\hat{x}(k+1|k) \in Y(k)$, we see that

$$P(k|N) + J(k)P_{\hat{x}\hat{x}}(k+1|N)J'(k) = P(k) + J(k)P_{\hat{x}\hat{x}}(k+1|k)J'(k),$$

or equivalently that

$$(53) \quad P(k|N) = P(k) + J(k)[P_{\hat{x}\hat{x}}(k+1|k) - P_{\hat{x}\hat{x}}(k+1|N)]J'(k).$$

But

$$\hat{x}(k+1|N) + \tilde{x}(k+1|N) = x(k+1),$$

and since $\hat{x}(k+1|N)$ and $\tilde{x}(k+1|N)$ are orthogonal, it follows that

$$P_{\hat{x}\hat{x}}(k+1|N) + P(k+1|N) = P_{xx}(k+1),$$

or that

$$(54) \quad P_{\hat{x}\hat{x}}(k+1|N) = P_{xx}(k+1) - P(k+1|N),$$

where

$$P_{xx}(k+1) \triangleq E[x(k+1)x'(k+1)],$$

and

$$P(k+1|N) = E[\tilde{x}(k+1|N)\tilde{x}'(k+1|N)].$$

Similarly,

$$(55) \quad \begin{aligned} P_{\hat{x}\hat{x}}(k+1|k) &= P_{xx}(k+1) - P(k+1|k) \\ &= P_{xx}(k+1) - M(k+1), \end{aligned}$$

where

$$P(k+1|k) = E[\tilde{x}(k+1|k)\tilde{x}'(k+1|k)] \triangleq M(k+1).$$

Substituting (54) and (55) into (53), we obtain

$$(56) \quad P(k|N) = P(k) + J(k)[P(k+1|N) - M(k+1)]J'(k),$$

which is the desired result. We observe that (56) is an $n \times n$ first-order matrix difference equation for which the indexing is $k = N-1, N-2, \dots, 0$, i.e., computation of the smoothing error covariance matrix proceeds backward in time. The information needed from the prediction and filtering solution in order to compute $P(k|N)$ obviously consists of the prediction and filtering error covariance matrices for $k = N-1, \dots, 0$.

Equation (56) gives the smoothing error covariance matrix for the smoothing formulation given by (46) and (47). We now apply the above procedure to obtain the smoothing error covariance matrix equation for the second formulation. First, from the definition of estimation error and (49), we obtain

$$(57) \quad \begin{aligned} \tilde{x}(k|N) + K(N-1, k)\hat{x}(N) \\ = \tilde{x}(k|N-1) + K(N-1, k)\hat{x}(N|N-1). \end{aligned}$$

Noting again that $\tilde{x}(k|N)$ is orthogonal to $Y(N)$, $K(N-1, k)\hat{x}(N) \in Y(N)$, $\tilde{x}(k|N-1)$ is orthogonal to $Y(N-1)$ and

$$K(N-1, k)\hat{x}(N|N-1) \in Y(N-1),$$

we obtain

$$(58) \quad \begin{aligned} P(k|N) = P(k|N-1) + K(N-1, k)[P_{\hat{x}\hat{x}}(N|N-1) \\ - P_{\hat{x}\hat{x}}(N)]K'(N-1, k). \end{aligned}$$

For $k = N-1$, (55) becomes

$$(59) \quad P_{\hat{x}\hat{x}}(N|N-1) = P_{xx}(N) - M(N).$$

From $\hat{x}(N) + \tilde{x}(N) = x(N)$, we obtain

$$(60) \quad P_{\hat{x}\hat{x}}(N) = P_{xx}(N) - P(N).$$

Substituting (59) and (60) into (58), we have now that

$$(61) \quad \begin{aligned} P(k|N) \\ = P(k|N-1) + K(N-1, k)[P(N) - M(N)]K'(N-1, k). \end{aligned}$$

Finally, one more simplification is possible in (61). Namely, we see from (11) for $k = N-1$ that

$$P(N) - M(N) = -K^0(N)H(N)M(N),$$

and we can write (61) as

$$(62) \quad \begin{aligned} & P(k|N) \\ &= P(k|N-1) - K(N-1, k)K^0(N)H(N)M(N)K'(N-1, k). \end{aligned}$$

This completes our development of the smoothing error covariance matrix equations for the two smoothing formulations. We remark that (56) and (62) agree with those obtained by Rauch, Tung, and Striebel [7] and by Rauch [16], respectively.

6. Summary of results. We summarize our results in two theorems.

THEOREM 4. *The optimal smoothed estimate of $x(k)$ for N measurements, $z(1), \dots, z(N)$, $k < N$, is given by the first-order system of linear difference equations*

$$(46) \quad \hat{x}(k|N) = \hat{x}(k) + J(k)[\hat{x}(k+1|N) - \hat{x}(k+1|k)],$$

where

$$(47) \quad J(k) = P(k)\Phi'(k+1, k)M^{-1}(k+1)$$

for $k = N-1, \dots, 0$. The corresponding smoothing error covariance matrix is given by the first-order system of matrix linear difference equations

$$(56) \quad P(k|N) = P(k) + J(k)[P(k+1|N) - M(k+1)]J'(k)$$

for $k = N-1, \dots, 0$.

THEOREM 5. *The optimal smoothed estimate of $x(k)$ for N measurements given the optimal smoothed estimate of $x(k)$ for $(N-1)$ measurements, $k < N$, is given by the first-order system of linear difference equations*

$$(49) \quad \hat{x}(k|N) = \hat{x}(k|N-1) + K(N-1, k)[\hat{x}(N) - \hat{x}(N|N-1)],$$

where

$$(50) \quad K(N-1, k) = \prod_{i=k}^{N-1} J(i)$$

and

$$(51) \quad J(i) = P(i)\Phi'(i+1, i)M^{-1}(i+1)$$

for $N = k+1, k+2, \dots$. The corresponding smoothing error covariance matrix is given by the first-order system of matrix linear difference equations

$$(62) \quad \begin{aligned} & P(k|N) = P(k|N-1) \\ & \quad - K(N-1, k)K^0(N)H(N)M(N)K'(N-1, k) \end{aligned}$$

for $N = k+1, k+2, \dots$.

7. Conclusion. We have presented here a derivation of the equations of discrete optimal linear smoothing for discrete linear systems using the method of orthogonal projection. The results are not new, but when combined with Kalman's results [1] on prediction and filtering give a unified treatment of the estimation problem in terms of orthogonal projection as remarked in §1.

Finally, we remark that a third formulation of the smoothing filter and smoothing error covariance equations, due originally to Rauch [16], can also be obtained using the present approach.

REFERENCES

- [1] R. E. KALMAN, *A new approach to linear filtering and prediction problems*, Trans. ASME Ser. D. J. Basic Engrg., 82D (1960), pp. 35-45.
- [2] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1955.
- [3] V. S. PUGACHEV, *On a possible general solution of the problem of determining optimum dynamic systems*, Avtomat. i Telemekh., 17 (1956), pp. 585-589.
- [4] R. E. KALMAN AND R. S. BUCY, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83D (1961), pp. 95-107.
- [5] Y. C. HO, *The method of least squares and optimal filtering theory*, RM-3329-PR, The Rand Corporation, Santa Monica, California 1962.
- [6] R. H. BATTIN, *Astronautical Guidance*, McGraw-Hill, New York, 1964, pp. 303-340.
- [7] H. E. RAUCH, F. TUNG AND C. T. STRIEBEL, *Maximum likelihood estimates of linear dynamic systems*, AIAA J., 3 (1965), pp. 1445-1450.
- [8] S. L. FAGIN, *Recursive linear regression theory, optimal filter theory, and error analyses of optimal systems*, 1964 IEEE International Convention Record, Part I, New York, 1964, pp. 216-240.
- [9] H. COX, *On the estimation of state variables and parameters for noisy dynamic systems*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 5-12.
- [10] Y. C. HO, *On the stochastic approximation method and optimal filtering theory*, J. Math. Anal. Appl., 6 (1962), pp. 152-154.
- [11] Y. C. HO AND R. C. K. LEE, *A Bayesian approach to problems in stochastic estimation and control*, Joint Automatic Control Conference, Stanford, 1964.
- [12] A. E. BRYSON AND M. FRAZIER, *Smoothing for linear and nonlinear dynamic systems*, TDR-63-119, Aeronautical Systems Division, Wright-Patterson AFB, Dayton, 1963.
- [13] R. E. KALMAN, *New methods and results in linear prediction and filtering theory*, Technical Report 61-1, RIAS, Baltimore, 1961.
- [14] P. D. JOSEPH, *Automatic rendezvous. Part II: Onboard navigation for rendezvous missions*, UCLA Engineering Extension Short Course: Space Control Systems—Attitude Control, Rendezvous, and Docking, Los Angeles, 1964.
- [15] B. FRIEDMAN, *Principles and Techniques of Applied Mathematics*, John Wiley, New York, 1956, Chap. 1.
- [16] H. E. RAUCH, *Solutions to the smoothing problem*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 371-372.

AN ABSTRACT VARIATIONAL THEORY WITH APPLICATIONS TO A BROAD CLASS OF OPTIMIZATION PROBLEMS. II.

APPLICATIONS*

LUCIEN W. NEUSTADT†

Abstract. In this article the general necessary conditions for variational problems obtained in Part I are specialized to a number of particular problems. These problems include all of the important optimal control problems, both with and without restricted phase coordinates, problems arising from quasiconvex families of functions, and discrete optimal control problems of various types. The necessary conditions obtained include, as special cases, the Pontryagin maximum principle and its various extensions and generalizations, as well as the associated transversality conditions.

1. Introduction. In this article we shall specialize the general necessary conditions obtained in [1] to a number of particular variational problems which include some general optimal control problems.

In §2 we define a locally convex linear topological function space \mathcal{S} and subspaces \mathcal{C} and \mathcal{S}_0 of \mathcal{S} . Then we define certain sets in \mathcal{S} (and \mathcal{S}_0) which represent the solutions of certain families of differential equations that arise in the theory of optimal control. Further, we construct first-order, convex approximations to these sets (for definitions of the terminology, see [1]). In §3 we construct a first-order, convex approximation to a set of functions in \mathcal{C} which represents the solutions of a collection of differential equations whose "right-hand sides" form a *quasiconvex* family of functions. Section 4 is devoted to finding a first-order, convex approximation to a set in a finite-dimensional space which arises from a class of difference equations. Some functionals which are differentiable in the sense of [1, (4.3) or (4.6)] are presented in §5, and their differentials are obtained.

In §§6–9 we consider a number of variational problems (which include most of the important optimal control problems) which are special cases of the canonical optimization problem defined in [1, §4], and, on the basis of [1, Theorems 4.1 and 4.2] and the results of §§2–5, obtain necessary conditions which solutions of these problems must satisfy. For the conventional optimal control problem, these conditions are equivalent to the Pontryagin maximum principle and the associated transversality conditions (see [2, Chap. I]), which, as is known, imply all of the first-order necessary conditions in the classical calculus of variations (see [2, Chap. V]).

A central theme in this work is the concept of quasiconvexity first introduced by Gamkrelidze in [3].

* Received by the editors June 1, 1966.

† Department of Electrical Engineering, University of Southern California, Los Angeles, California 90007. This work was supported by the United States Air Force Office of Scientific Research under Grant AF-AFOSR-1029-66.

2. Some particular first-order, convex approximations. In this section we shall construct first-order convex approximations to certain sets, which are important in applications, in a particular locally convex, linear topological space. For the definition of first-order, convex approximations, see [1, §2].

Let $I = [t_1, t_2]$ be a fixed compact interval, and let \mathcal{C} denote the linear space of continuous functions from I into R^n (Euclidean n -space), where n is a fixed positive integer, normed by means of the relation

$$(2.1) \quad \|x\| = \sup_{t \in I} |x(t)|,$$

where $|\xi|$ denotes the Euclidean length of the vector $\xi \in R^n$. (This notation will be used throughout in the sequel.) The components of a vector $\xi \in R^n$ will be denoted by $\xi^j, j = 1, \dots, n$.

Let \mathcal{S}_1 denote the set of all functions x from I to R^n for which there exists a sequence of functions $x_i, i = 1, 2, \dots$, such that (1) $x_i \in \mathcal{C}$ for each i , (2) $x_{i+1}^j(t) \geq x_i^j(t)$ for all $t \in I, j = 1, \dots, n, i = 1, 2, \dots$, and (3) $\lim_{i \rightarrow \infty} x_i(t) = x(t)$ for every $t \in I$. Let $\mathcal{S} = \{x: x = x' - x'', x' \in \mathcal{S}_1, x'' \in \mathcal{S}_1\}$. It is evident that \mathcal{S} is a linear vector space, that $\mathcal{C} \subset \mathcal{S}$, and that every piecewise-continuous function from I to R^n , all of whose discontinuities are of the first kind, belongs to \mathcal{S} . Let us define a topology on \mathcal{S} by taking as a base all sets of the form $\{y: |y(s_i) - x(s_i)| < \epsilon, i = 1, \dots, \nu\}$, where $\{s_1, \dots, s_\nu\}$ is an arbitrary finite subset of I, x is an arbitrary element of \mathcal{S} , and ϵ is an arbitrary positive number. It is easily seen that \mathcal{S} is a locally convex, linear topological space under this topology. Further (see [4, p. 421, Theorem 9]), the conjugate space \mathcal{S}^* of \mathcal{S} consists of all functionals of the form $l(x) = \sum_{i=1}^\nu \alpha_i x(s_i)$, where the α_i are arbitrary real numbers, s_1, \dots, s_ν are arbitrary points in I , and ν is an arbitrary positive integer.

Let $\mathcal{S}_0 = \{x: x \in \mathcal{S}, x(t_1) = 0\}$, so that \mathcal{S}_0 is a closed subspace of \mathcal{S} . We shall consider \mathcal{S}_0 to be a linear topological space by letting the topology on \mathcal{S}_0 be the one induced by the topology on \mathcal{S} .

Let $f(x, u)$ be a given function from $G \times U$ into R^n , where G is a non-empty open set in R^n and U is an arbitrary fixed set in R^r containing more than one point. Let us suppose that f is continuous in both x and u , and is of class C^1 with respect to $x \in G$. Let Ω denote the set of all measurable, essentially bounded functions from I to U , and let Q_0' denote the set of all absolutely continuous functions $x \in \mathcal{C}$ whose range is contained in G which, for some function $u \in \Omega$, satisfy the equation

$$(2.2) \quad \dot{x}(t) = f(x(t), u(t))$$

for almost all $t \in I$. Finally, let z be a fixed element of Q_0' , and let $Q_0 = Q_0' - z$. Thus, $0 \in Q_0$ and $Q_0 \subset \mathcal{S}$. Since $z \in Q_0'$, there is a function

$u^* \in \Omega$ such that

$$(2.3) \quad \dot{z}(t) = f(z(t), u^*(t)) \quad \text{for almost all } t \in I.$$

Let I_1 denote the set of points $t \in (t_1, t_2)$ for which $u^*(t)$ is regular, i.e., $t \in I_1$ if and only if $t_1 < t < t_2$ and the relation

$$(2.4) \quad \lim_{\text{meas } J \rightarrow 0} \frac{\text{meas } [u^{*-1}(O) \cap J]}{\text{meas } J} = 1$$

holds for every neighborhood O of $u^*(t)$. In (2.4), J is an arbitrary subinterval of I such that $t \in J$. It is well known that $\text{meas } (I_1) = \text{meas } (I)$.

Let \mathcal{C}_0 denote the finite-dimensional subspace of \mathcal{C} consisting of all absolutely continuous functions $x \in \mathcal{C}$ that satisfy the linear equation

$$(2.5) \quad \dot{x}(t) = f_x(z(t), u^*(t))x(t) \quad \text{for almost all } t \in I,$$

where f_x denotes the matrix whose elements are the partial derivatives $\partial f^i / \partial x^j$. For every $s \in I_1$, $v \in U$, and $\xi \in R^n$, let $y_{s,v}(t)$ and $y_\xi(t)$ be the functions in \mathcal{C}_0 that satisfy the initial conditions

$$(2.6) \quad y_{s,v}(s) = f(z(s), v) - f(z(s), u^*(s)), \quad y_\xi(t_1) = \xi.$$

Finally, for every $(s, v, \xi) \in I_1 \times U \times R^n$, let $x_{s,v,\xi} \in \mathcal{S}$ be defined as follows:

$$(2.7) \quad x_{s,v,\xi}(t) = \begin{cases} y_\xi(t) & \text{for } t_1 \leq t < s, \\ y_{s,v}(t) + y_\xi(t) & \text{for } s \leq t \leq t_2, \end{cases}$$

and let

$$K_1 = \left\{ \sum_{i=1}^{\nu} \beta_i x_{s_i, v_i, \xi_i} : s_i \in I_1, v_i \in U, \xi_i \in R^n, \beta_i \geq 0 \right. \\ \left. \text{for } i = 1, \dots, \nu; \nu \text{ arbitrary} \right\}.$$

It is clear that K_1 is a convex cone in \mathcal{S} with vertex at 0. It follows from the results in [2, Chap. II, and in particular §§13 and 14, pp. 86–99] that K_1 is a first-order, convex approximation to Q_0 , where the underlying space \mathfrak{S} is \mathcal{S} . Indeed, K_1 is closely related to the cone of attainability described in [2, Chap. II]. It also follows that the convex cone

$$(2.8) \quad K_0 = \left\{ \sum_{i=1}^{\nu} \beta_i x_{s_i, v_i, 0} : s_i \in I_1, v_i \in U, \beta_i \geq 0 \right. \\ \left. \text{for } i = 1, \dots, \nu; \nu \text{ arbitrary} \right\}$$

is a first-order, convex approximation to $Q_0 \cap \mathcal{S}_0$, where the underlying space \mathfrak{S} is \mathcal{S}_0 .

If the function $f(x, u)$ is of class C^1 with respect to both x and u , and if $u^*(t) \in (\text{interior of } U)$ for each $t \in I$, we can obtain another first-order, convex approximation to Q_0 (and to $Q_0 \cap S_0$). Namely, for every $u \in L_r^\infty(I)$ (i.e., for every measurable, essentially bounded function u from I to R^n), let y_u denote the absolutely continuous function in \mathcal{C} that satisfies the relations

$$(2.9) \quad \frac{dy_u(t)}{dt} = f_x(z(t), u^*(t))y_u(t) + f_u(z(t), u^*(t))u(t)$$

for almost all $t \in I$, $y_u(t_1) = 0$,

where f_u denotes the matrix whose elements are the partial derivatives $\partial f^i / \partial u^j$. Let $K_0' = \{y_u : u \in L_r^\infty(I)\}$, $K_1' = \{y_u + y_\xi : u \in L_r^\infty(I), \xi \in R^n\}$. It is easily seen that K_0' and K_1' are convex cones in $\mathcal{C} \subset S$ with vertex at 0, and it follows from standard arguments, which are similar to those used to develop the dependence of solutions of differential equations on parameters, that K_1' and K_0' are first-order, convex approximations to Q_0 and $Q_0 \cap S_0$, respectively, where the underlying spaces are \mathcal{C} and $\mathcal{C} \cap S_0$, respectively. (Similar arguments are used to prove Theorem 3.1 in the sequel.)

Let us consider a subset Ω_1 of Ω and a corresponding subset Q_1' of Q_1 . Namely, let Ω_1 be the set of all functions in Ω that are piecewise-constant, with their points of discontinuity all belonging to a preassigned, finite subset $I_0 = \{s_1^*, \dots, s_\mu^*\}$ of (t_1, t_2) . Let Q_1' be defined in terms of Ω_1 in the same way that Q_0' was defined in terms of Ω , let $z \in Q_1'$ so that (2.3) holds with some $u^* \in \Omega_1$, and let $Q_1 = Q_1' - z$. We shall again assume that $u^*(t) \in (\text{interior of } U)$ for every $t \in I$, and that f_u exists and is continuous in $G \times U$. Then if

$$(2.10) \quad K_0'' = \{y_u : u \in L_r^\infty(I), u \text{ piecewise-constant with points of discontinuity in } I_0\},$$

$$K_1'' = \{y_u + y_\xi : y_u \in K_0'', \xi \in R^n\}$$

(where y_u and y_ξ are again defined as solutions of (2.9) and (2.5) (with (2.6)), respectively), it is easily seen that K_1'' and K_0'' are first-order, convex approximations to Q_1 and $(Q_1 \cap S_0)$, respectively, where the underlying spaces are \mathcal{C} and $\mathcal{C} \cap S_0$, respectively.

As a similar example, consider the case where Ω_2 is the set of all functions in Ω which are piecewise-constant, and whose points of discontinuity belong to (t_1, t_2) and are not more than μ in number, where μ is a pre-assigned positive integer. Define Q_2' in terms of Ω_2 in the same way that Q_0' was defined in terms of Ω , let $z \in Q_2'$ so that (2.3) holds with some $u^* \in \Omega_2$, and let $Q_2 = Q_2' - z$. We shall again suppose that $u^*(t) \in (\text{interior of } U)$ for every $t \in I$, and that f is of class C^1 with respect to

x and u . Let $I^* = \{s_1^*, \dots, s_{\bar{\mu}}^*\}$ be the set of points in (t_1, t_2) at which u^* is discontinuous, $\bar{\mu} \leq \mu$. Let y_i^* for $i = 1, \dots, \bar{\mu}$ denote the absolutely continuous function in \mathfrak{C} that satisfies (2.5) together with the initial condition $y_i^*(s_i^*) = f(z(s_i^*), u^*(s_i^{*+})) - f(z(s_i^*), u^*(s_i^{*-}))$, and let y_i^+ and y_i^- , $i = 1, \dots, \bar{\mu}$, denote the functions in \mathfrak{S} defined by the relations

$$y_i^+(t) = \begin{cases} 0 & \text{for } t_1 \leq t < s_i^*, \\ y_i^*(t) & \text{for } s_i^* \leq t \leq t_2, \end{cases}$$

$$y_i^-(t) = \begin{cases} 0 & \text{for } t_1 \leq t \leq s_i^*, \\ y_i^*(t) & \text{for } s_i^* < t \leq t_2. \end{cases}$$

Then, if

$$(2.11) \quad K_0''' = \left\{ y_u + \sum_{i=1}^{\bar{\mu}} \eta_i y_i^{\text{sgn } \eta_i} : u \in L_r^\infty(I) \text{ and piecewise-constant} \right. \\ \left. \text{with points of discontinuity in } I^*, (\eta_1, \dots, \eta_{\bar{\mu}}) \in R^{\bar{\mu}} \right\},$$

$$K_1'' = \{y + y_i : y \in K_0''', \xi \in R^n\},$$

where y_u is given by (2.9) and y_ξ by (2.5) and (2.6), it is not difficult to show that K_1''' and K_0''' are first-order, convex approximations to Q_2 and $(Q_2 \cap \mathfrak{S}_0)$, respectively, where the underlying spaces are \mathfrak{S} and \mathfrak{S}_0 , respectively.

Note 2.1. If in the two examples described in the preceding two paragraphs, the words "piecewise-constant" are replaced by "piecewise-continuous with discontinuities of the first kind", all of the conclusions are easily seen to remain in force.

Note 2.2. In the two preceding examples, as well as in the development of K_0' and K_1' , the requirement that $u^*(t) \in (\text{interior of } U)$ can be replaced by the weaker requirement that U , in a neighborhood of $u^*(t)$, can, for each $t \in I$, be approximated in a certain sense (which we shall not here describe in detail) by a convex set. Then, if K_0' , K_1' , etc., are suitably re-defined, they will be first-order, convex approximations to Q_0 , $Q_0 \cap \mathfrak{S}_0$, etc.

Note 2.3. The last two examples can be generalized as follows. Let ρ_1, \dots, ρ_l , where $l \leq \bar{\mu}$, and $\sigma_1, \dots, \sigma_k$ be preassigned continuously differentiable functions from R^μ to R^1 . Let Ω_3 be the set of all functions $u \in \Omega_2$ with the property that there exist points s_1, \dots, s_μ (which in general depend on u) which (a) include all of the points of discontinuity of u , (b) all belong to (t_1, t_2) , (c) satisfy the equations $\rho_j(s_1, \dots, s_\mu) = 0$ for $j = 1, \dots, l$, and (d) satisfy the inequalities $\sigma_j(s_1, \dots, s_\mu) \leq 0$ for $j = 1, \dots, k$. We can correspondingly define the set Q_3' . Then, for an element $z \in Q_3'$, we can, in a manner similar to that used to define K_1''' and K_0''' , obtain first-order, convex approximations to $(Q_3' - z)$ and

$(Q_3' - z) \cap S_0$. Here it is again necessary to assume that f_u exists and is continuous, and that $u^*(t) \in (\text{interior of } U)$ for every $t \in I$, or that U , near $u^*(t)$, is as indicated in Note 2.2. The remark in Note 2.1 is also pertinent for this example.

3. Quasiconvexity and first-order, convex approximations. In this section we shall construct a first-order, convex approximation in the space \mathcal{C} to a set of functions (in \mathcal{C}) which constitute the solutions of a collection of differential equations whose "right-hand sides" generate a quasiconvex family of functions. The concept of quasiconvexity, which was introduced in [3], appears to be particularly suitable for studying a wide class of variational problems.

We shall begin with some definitions.

Let I again be a fixed compact interval $[t_1, t_2]$, and let G be an open set in R^n . We shall denote by $L^1(I)$ the set of all real-valued functions defined on I and integrable over I , and by $L^\infty(I)$ the set of all real-valued, measurable, essentially bounded functions defined on I .

Let \mathcal{G} denote the linear vector space of all functions $g(x, t)$ from $G \times I$ into R^n which are of class C^1 with respect to $x \in G$, measurable with respect to $t \in I$, and are such that $|g(x, t)| \in L^1(I)$ for every fixed $x \in G$.

For every subset X of G , we shall define a corresponding topology on \mathcal{G} , which we shall refer to as the X -topology. Namely, the X -topology of \mathcal{G} is the topology obtained by taking as base all sets of the form

$$\left\{ g: g \in \mathcal{G}, \left| \int_{t_1}^{t_2} [g(x_i, t) - g_0(x_i, t)] \theta_j(t) dt \right| < \epsilon, \right. \\ \left. i = 1, \dots, l, \quad j = 1, \dots, k \right\},$$

where $\{x_1, \dots, x_l\}$ is a finite subset of X , $\{\theta_1, \dots, \theta_k\}$ is a finite subset of $L^\infty(I)$, $\epsilon > 0$, and $g_0 \in \mathcal{G}$.

The X -topology on \mathcal{G} is a natural analog of the weak topology on $L^1(I)$.

DEFINITION 3.1. If X is a subset of G , we shall say that a set $D \subset \mathcal{G}$ is *dominated over X* if there exists a function $\bar{m}(t) \in L^1(I)$ such that $|g(x, t)| + |g_x(x, t)| \leq \bar{m}(t)$ for all $(x, t) \in X \times I$ and $g \in D$.

Throughout the sequel, for every positive integer ν , we shall denote by P^ν the following subset of $R^\nu: P^\nu = \{\beta = (\beta_1, \dots, \beta_\nu): \beta_i \geq 0 \text{ for } i = 1, \dots, \nu, \sum_{i=1}^\nu \beta_i = 1\}$.

DEFINITION 3.2. A set Γ of functions $g(x, t)$ from $G \times I$ into R^n will be said to be *quasiconvex* if $\Gamma \subset \mathcal{G}$ and if, for every finite subset $\{g_0, \dots, g_m\}$ of Γ and every compact subset X of G , there exists a set $D \subset \mathcal{G}$ dominated over X such that, for every neighborhood N of 0 in the X -topology of \mathcal{G} ,

there exists a map γ (which may depend on X , the g_i , and N) from P^{m+1} to $(\Gamma \cap D)$, continuous in the X -topology, such that $[\gamma(\beta_0, \dots, \beta_m) - \sum_{i=0}^m \beta_i g_i] \in (N \cap D)$ for every $\beta = (\beta_0, \dots, \beta_m) \in P^{m+1}$.

Definition 3.2 differs slightly from the definition of quasiconvexity given in [3, Definition 2.2]. It is not difficult to verify that a set quasiconvex in the sense of [3] is also quasiconvex in the sense of our Definition 3.2. Definition 3.2 is based on a suggestion of H. Hermes. Gamkrelidze also suggested a definition somewhat at variance with his original one, which is very similar to, but not quite the same, as the one given here.

Now let Γ be a given quasiconvex set of functions in \mathcal{G} . Let $Q^{*'} denote the set of all absolutely continuous functions $x \in \mathcal{C}$ (where \mathcal{C} is as defined in §2), whose range is contained in G , and which, for some function $g \in \Gamma$, satisfy the equation$

$$\dot{x}(t) = g(x(t), t) \quad \text{for almost all } t \in I.$$

Let z be a fixed element of $Q^{*'}$, so that

$$(3.1) \quad \dot{z}(t) = g^*(z(t), t) \quad \text{for almost all } t \in I, \quad \text{where } g^* \in \Gamma,$$

and let $Q^* = Q^{*' - z}$. Thus, $0 \in Q^* \subset \mathcal{C}$.

We shall construct a first-order, convex approximation to Q^* . Many of the arguments below are patterned after the proof of Theorem 2.1 in [3, §3].

Let $[\Gamma]$ denote the convex hull of Γ , i.e.,

$$[\Gamma] = \left\{ \sum_{i=1}^{\nu} \beta_i g_i(x, t) : (\beta_1, \dots, \beta_{\nu}) \in P^{\nu}, g_i \in \Gamma \text{ for each } i, \right. \\ \left. \nu \text{ an arbitrary positive integer} \right\}.$$

For each $\xi \in R^n$ and $h \in [\Gamma]$, let $\delta x_{\xi, h}$ denote the following function in \mathcal{C} :

$$(3.2) \quad \delta x_{\xi, h}(t) = \Phi(t) \left[\xi + \int_{t_1}^t \Phi^{-1}(\tau) [h(z(\tau), \tau) - g^*(z(\tau), \tau)] d\tau \right] \\ \text{for each } t \in I,$$

where $\Phi(t)$ is the absolutely continuous $n \times n$ matrix-valued function defined on I that satisfies the relations

$$(3.3) \quad \dot{\Phi}(t) = g_x^*(z(t), t) \Phi(t) \quad \text{for almost all } t \in I, \\ \Phi(t_1) = \text{the identity matrix},$$

where g_x is the $n \times n$ Jacobian matrix derived from g . Let

$$(3.4) \quad K^* = \{\delta x_{\xi, h} : \xi \in R^n, h \in [\Gamma]\}.$$

THEOREM 3.1. *Let Γ be a quasiconvex subset of \mathfrak{g} , and let $Q^{*'}$ be defined as above. Then if $z \in Q^{*'}$, the set K^* defined as above (see (3.1)–(3.4)) is a first-order, convex approximation to $Q^* = Q^{*'} - z$ in the space \mathcal{C} .*

Proof. Note that $\delta x_{0, \mathfrak{g}^*} = 0 \in \mathcal{C}$, so that $0 \in K^*$. It follows at once from (3.2) that, for every $\beta = (\beta_1, \dots, \beta_\nu) \in P^\nu$ (where ν is an arbitrary positive integer),

$$(3.5) \quad \sum_{i=1}^{\nu} \beta_i \delta x_{\xi_i, h_i} = \delta x_{\sum \beta_i \xi_i, \sum \beta_i h_i} \in K^*,$$

so that K^* is convex.

Let $\{\delta x_{\xi_1, h_1}, \dots, \delta x_{\xi_\nu, h_\nu}\}$ be an arbitrary finite subset of K^* , which will remain fixed for the remainder of the argument. Now $h_j \in [\Gamma]$ for $j = 1, \dots, \nu$; therefore, there are functions $g_i(x, t) \in \Gamma$, $i = 1, \dots, m$, and vectors $(\beta_{1j}, \dots, \beta_{mj}) \in P^m$, $j = 1, \dots, \nu$, such that

$$h_j(x, t) = \sum_{i=1}^m \beta_{ij} g_i(x, t) \quad \text{for } j = 1, \dots, \nu.$$

For every $\beta = (\beta_1, \dots, \beta_\nu) \in P^\nu$, let us denote $\sum_{i=1}^{\nu} \beta_i \delta x_{\xi_i, h_i}(t)$ by $\delta x(t; \beta)$, and $\sum_{i=1}^{\nu} \beta_i [h_i(x, t) - g^*(x, t)]$ by $\delta g(x, t; \beta)$. It now follows at once that, if $0 \leq \epsilon \leq 1$ and $(\beta_1, \dots, \beta_\nu) \in P^\nu$, then

$$(3.6) \quad g^*(x, t) + \epsilon \delta g(x, t; \beta) = (1 - \epsilon) g^*(x, t) + \sum_{i=1}^m (\epsilon \bar{\beta}_i) g_i(x, t)$$

and $(1 - \epsilon, \epsilon \bar{\beta}_1, \dots, \epsilon \bar{\beta}_m) \in P^{m+1}$,

where $\bar{\beta}_i = \sum_{j=1}^{\nu} \beta_j \beta_{ij}$ for $i = 1, \dots, m$. Further, it is an immediate consequence of (3.2), (3.3) and (3.5), that, for every $\beta = (\beta_1, \dots, \beta_\nu) \in P^\nu$,

$$(3.7) \quad \frac{d}{dt} \delta x(t; \beta) = g_x^*(z(t), t) \delta x(t; \beta) + \delta g(z(t), t; \beta)$$

for almost all $t \in I$,

$$\delta x(t_1; \beta) = \sum_{i=1}^{\nu} \beta_i \xi_i.$$

Let X_0 and X_1 be compact subsets of G such that $z(t) \in (\text{interior of } X_1)$ for every $t \in I$, and $X_1 \subset (\text{interior of } X_0)$. Such sets exist because G is open and $\{z(t) : t \in I\}$ is a compact subset of G . Let η_0 be the distance from X_1 to the complement of X_0 ; $\eta_0 > 0$.

We now prove the following lemma.

LEMMA 3.1. *Let $D_1 \subset \mathfrak{g}$ be dominated over X_0 and let $m^*(t) \in L^1(I)$ (in other respects, D_1 and m^* are arbitrary). Then, for every $\eta > 0$, there is a neighborhood N_η (in the X_1 topology of \mathfrak{g} , and possibly depending on η , D_1 ,*

and m^*) of 0 in \mathcal{G} such that

$$\left| \int_{t_1}^{\tau} g(y(t), t) dt \right| < \eta \quad \text{for all } \tau \in I,$$

for every $g \in (N_{\eta} \cap D_1)$, and for all absolutely continuous functions $y(t)$ from I into X_1 that satisfy the inequality $|\dot{y}(t)| \leq m^*(t)$ for almost all $t \in I$.

Proof. Since D_1 is dominated over X_0 , there is a function $\tilde{m}(t) \in L^1(I)$ such that $|g(x, t)| + |g_x(x, t)| \leq \tilde{m}(t)$ for all $(x, t) \in X_0 \times I$ and $g \in D_1$.

Let $\theta = \min \{ \eta_0, \eta(2 + 4 \int_I \tilde{m}(t) dt)^{-1} \}$, where $\eta > 0$ is preassigned, and let s_0, s_1, \dots, s_k be points in I such that $s_0 = t_1, s_k = t_2, s_j < s_{j+1}$ for $j = 0, 1, \dots, k-1$, and such that

$$(3.8) \quad \int_{s_{j-1}}^{s_j} [m^*(t) + \tilde{m}(t)] dt < \theta \quad \text{for } j = 1, \dots, k.$$

Since X_1 is compact, there is a finite subset $\{\chi_1, \dots, \chi_l\}$ of X_1 such that, for every $\chi \in X_1$, there is a point $\chi_i, 1 \leq i \leq l$, with $|\chi - \chi_i| < \theta$. By definition of θ and η_0 , this means that the entire line segment joining χ and χ_i is contained in X_0 . Let

$$(3.9) \quad N_{\eta} = \left\{ g: g \in \mathcal{G}, \left| \int_{s_{j-1}}^{s_j} g(\chi_i, t) dt \right| < \frac{\eta}{2k} \right. \\ \left. \text{for } i = 1, \dots, l \text{ and } j = 1, \dots, k \right\}.$$

Now let $\tau \in I$, and let $y(t) \in \mathcal{C}$ satisfy the hypotheses in the statement of the lemma. For each $j = 0, 1, \dots, k$ there is an integer $i_j, 1 \leq i_j \leq l$, such that

$$(3.10) \quad |y(s_j) - \chi_{i_j}| < \theta,$$

and it follows from (3.8) that, whenever $t \in [s_{j-1}, s_j]$,

$$(3.11) \quad |y(t) - y(s_j)| < \theta,$$

and the entire line segment joining $y(t)$ with $y(s_j)$ is contained in X_0 . Let j' be such that $0 \leq j' \leq k-1$ and $\tau \in [s_{j'}, s_{j'+1}]$. It now follows from (3.8)–(3.11) that if $g \in N_{\eta} \cap D_1$, then

$$\left| \int_{t_1}^{\tau} g(y(t), t) dt \right| = \left| \sum_{j=1}^{j'} \int_{s_{j-1}}^{s_j} g(y(t), t) dt + \int_{s_{j'}}^{\tau} g(y(t), t) dt \right| \\ \leq \sum_{j=1}^{j'} \left\{ \left| \int_{s_{j-1}}^{s_j} g(\chi_{i_j}, t) dt \right| + \int_{s_{j-1}}^{s_j} [|g(y(s_j), t) - g(\chi_{i_j}, t)| \right.$$

$$\begin{aligned}
 & + |g(y(t), t) - g(y(s_j), t)| \, dt \Big\} + \int_{s_j'}^{s_{j'+1}} \tilde{m}(t) \, dt \\
 & < \frac{j'\eta}{2k} + \sum_{j=1}^{j'} \int_{s_{j-1}}^{s_j} [\max_{x \in X_0} |g_x(x, t)| \\
 & \quad \cdot (|y(s_j) - \chi_{i_j}| + |y(t) - y(s_j)|) \, dt + \theta \\
 & \leq \frac{\eta}{2} + \theta \left(1 + 2 \int_I \tilde{m}(t) \, dt \right) \leq \eta.
 \end{aligned}$$

This completes the proof of Lemma 3.1.

Since g^*, g_1, \dots, g_m all belong to Γ , it follows from (3.6), Definition 3.2 and Lemma 3.1 that there exist a function $m_0(t) \in L^1(I)$ and functions $g(x, t; \beta, \epsilon) \in \Gamma$, defined for every $\beta = (\beta_1, \dots, \beta_\nu) \in P^\nu$ and $\epsilon \in [0, 1]$, such that if

$$(3.12) \quad \delta^2 g(x, t; \beta, \epsilon) = g(x, t; \beta, \epsilon) - [g^*(x, t) + \epsilon \delta g(x, t; \beta)],$$

then

$$(3.13) \quad |g(x, t; \beta, \epsilon)| + |g_x(x, t; \beta, \epsilon)| + |\delta^2 g(x, t; \beta, \epsilon)| \\ + |\delta^2 g_x(x, t; \beta, \epsilon)| \leq m_0(t) \quad \text{for all } (x, t) \in X_0 \times I,$$

$$(3.14) \quad \left| \int_{t_1}^\tau \delta^2 g(y(t), t; \beta, \epsilon) \, dt \right| < \epsilon^2 \quad \text{for all } \tau \in I$$

and for every absolutely continuous function $y(t)$ from I into X_1 that satisfies the inequality $|y(t)| \leq m_0(t)$ for almost all $t \in I$, and

$$(3.15) \quad g(x, t; \beta', \epsilon) \xrightarrow[\beta' \in P^\nu]{\beta' \rightarrow \beta} g(x, t; \beta, \epsilon) \quad \text{in the } X_0 \text{ topology of } \mathcal{G},$$

(relations (3.13)–(3.15) hold for all $\beta \in P^\nu$ and $\epsilon \in [0, 1]$).

Now consider the differential equations (see (3.12))

$$(3.16) \quad \dot{x}(t) = g(x(t), t; \beta, \epsilon) \\ = g^*(x(t), t) + \epsilon \delta g(x(t), t; \beta) + \delta^2 g(x(t), t; \beta, \epsilon),$$

where $\beta \in P^\nu$ and $0 \leq \epsilon \leq 1$. Let the solution of (3.16) with initial value $x(t_1) = z(t_1) + \epsilon \sum_{i=1}^\nu \beta_i \xi_i$, where $\beta = (\beta_1, \dots, \beta_\nu)$, be denoted by $x(t; \beta, \epsilon)$. We shall show that there is a number $\epsilon_1 > 0$ such that $x(t; \beta, \epsilon)$ is defined on all of I and takes on values in G for every $\beta \in P^\nu$ and $\epsilon \in [0, \epsilon_1]$ and that

$$(3.17) \quad \max_{t \in I} \left| \frac{x(t; \beta, \epsilon) - z(t)}{\epsilon} - \sum_{i=1}^\nu \beta_i \delta x_{\xi_i, h_i}(t) \right| \xrightarrow{\epsilon \rightarrow 0^+} 0$$

uniformly with respect to $\beta = (\beta_1, \dots, \beta_\nu) \in P^\nu$,

$$(3.18) \quad \max_{t \in I} |x(t; \beta', \epsilon) - x(t; \beta, \epsilon)| \xrightarrow[\beta' \in P^v]{\beta' \rightarrow \beta} 0$$

for every $\beta \in P^v$ and $\epsilon \in [0, \epsilon_1]$.

But this will imply that $[x(t; \beta, \epsilon) - z(t)] \in (Q^{**} - z) = Q^*$ for every $\beta \in P^v$ and $\epsilon \in [0, \epsilon_1]$ (see (3.16) and recall that $g(x, t; \beta, \epsilon) \in \Gamma$), and it will then follow from (3.17) and (3.18) that K^* is a first-order, convex approximation to Q^* in the space \mathcal{C} .

Let us begin by showing that there is a number $\epsilon_1 > 0$ such that $x(t; \beta, \epsilon)$ is defined on all of I and has range contained in G for every $\beta \in P^v$ and $\epsilon \in [0, \epsilon_1]$, and that, in addition,

$$(3.19) \quad \max_{t \in I} |x(t; \beta, \epsilon) - z(t)| \xrightarrow[\epsilon \rightarrow 0^+]{\epsilon \rightarrow 0^+} 0$$

uniformly with respect to $\beta \in P^v$.

Because $z(t_1) \in (\text{interior of } X_1)$, there is a number $\bar{\epsilon}$, $0 < \bar{\epsilon} < 1$, such that $|\epsilon \sum_{i=1}^v \beta_i \xi_i| < \eta_0/2$ and $[z(t_1) + \epsilon \sum_{i=1}^v \beta_i \xi_i] \in (\text{interior of } X_1) \subset (\text{interior of } G)$ for every $\beta \in P^v$ and $\epsilon \in [0, \bar{\epsilon}]$. Hence, for all $\beta \in P^v$, $x(t; \beta, \epsilon)$ is defined and belongs to X_1 for t in a neighborhood of t_1 whenever $0 \leq \epsilon \leq \bar{\epsilon}$. Thus, for every $\epsilon \in [0, \bar{\epsilon}]$, there is a number $t_\epsilon \in I$ such that, for every $\beta \in P^v$ and $t \in [t_1, t_\epsilon]$, $x(t; \beta, \epsilon)$ is defined and belongs to X_1 and $|x(t; \beta, \epsilon) - z(t)| < \eta_0/2$. For each $\epsilon \in [0, \bar{\epsilon}]$, let t_ϵ be the l.u.b. of all such numbers t_ϵ , and let us denote $[t_1, t_\epsilon]$ by I_ϵ (if $t_\epsilon = t_1$, I_ϵ consists of the single point t_1).

Hence, for all $\beta \in P^v$ and $\epsilon \in [0, \bar{\epsilon}]$, $x(t; \beta, \epsilon) \in X_1$ for all $t \in I_\epsilon$, and it follows from (3.16) and (3.13) that $|x(t; \beta, \epsilon)| \leq m_0(t)$ for almost all $t \in I_\epsilon$. Now let the functions $y(t; \beta, \epsilon)$ be defined as follows: $y(t; \beta, \epsilon) = x(t; \beta, \epsilon)$ for $t \in I_\epsilon$, $y(t; \beta, \epsilon) = x(t_\epsilon; \beta, \epsilon)$ for $t_\epsilon \leq t \leq t_2$. It is clear that $y(t; \beta, \epsilon)$ is an absolutely continuous function from I into X_1 and that $|\dot{y}(t; \beta, \epsilon)| \leq m_0(t)$ for almost all $t \in I$. Hence (see (3.14)), for all $t \in I_\epsilon$, $\beta \in P^v$, and $\epsilon \in [0, \bar{\epsilon}]$, we have

$$(3.20) \quad \left| \int_{t_1}^t \delta^2 g(x(s; \beta, \epsilon), s; \beta, \epsilon) ds \right| < \epsilon^2.$$

If $\epsilon \in [0, \bar{\epsilon}]$, it follows from (3.16) and (3.1) and the definition of $x(t; \beta, \epsilon)$ that

$$(3.21) \quad \begin{aligned} x(t; \beta, \epsilon) - z(t) &= \epsilon \left[\sum_{i=1}^v \beta_i \xi_i + \int_{t_1}^t \delta g(x(s; \beta, \epsilon), s; \beta) ds \right] \\ &+ \int_{t_1}^t [g^*(x(s; \beta, \epsilon), s) - g^*(z(s), s)] ds \end{aligned}$$

$$+ \int_{t_1}^t \delta^2 g(x(s; \beta, \epsilon), s; \beta, \epsilon) ds$$

for all $t \in I_\epsilon$ and $\beta \in P''$. Since (3.13) holds for both $\epsilon = 0$ and $\epsilon = 1$, we have (see (3.12)) that

$$(3.22) \quad |g_x^*(x, t)| \leq m_0(t), \quad |\delta g(x, t; \beta)| \leq 2m_0(t) \\ \text{for all } (x, t) \in X_0 \times I \text{ and } \beta \in P''.$$

It now follows from (3.20)–(3.22) that, if $\epsilon \in [0, \bar{\epsilon}]$, then, for all $\beta \in P''$ and $t \in I_\epsilon$,

$$|x(t; \beta, \epsilon) - z(t)| \leq \epsilon \left[\sum_{i=1}^r \beta_i |\xi_i| + \int_{t_1}^t 2m_0(s) ds \right] \\ + \int_{t_1}^t [\sup_{x \in X_0} |g_x^*(x, s)|] |x(s; \beta, \epsilon) - z(s)| ds + \epsilon^2 \\ \leq \epsilon \left[\sum_{i=1}^r |\xi_i| + 2 \int_I m_0(s) ds + 1 \right] \\ + \int_{t_1}^t m_0(s) |x(s; \beta, \epsilon) - z(s)| ds.$$

Let $\bar{\eta} = \sum_{i=1}^r |\xi_i| + 2 \int_I m_0(s) ds + 1$ and $\tilde{\eta} = \bar{\eta} e^{\bar{\eta}}$. It now follows from Gronwall's inequality that

$$(3.23) \quad |x(t; \beta, \epsilon) - z(t)| < \epsilon \tilde{\eta} \quad \text{for all } (t, \beta) \in I_\epsilon \times P'' \text{ and } \epsilon \in [0, \bar{\epsilon}].$$

Let us denote the compact subset $\{z(t) : t \in I\}$ of G by Z_I . Let η_1 denote the distance from Z_I to the complement of X_1 . Since $Z_I \subset (\text{interior of } X_1)$, $\eta_1 > 0$. Let $\eta_2 = \min \{\eta_1, \eta_0/2\}$, $\epsilon_1 = \eta_2/(2\tilde{\eta})$. We shall now show that if $0 \leq \epsilon \leq \epsilon_1$, then $t_\epsilon = t_2$, or $I_\epsilon = I$. Indeed, suppose the contrary. Now, $|x(t_\epsilon; \beta, \epsilon) - z(t_\epsilon)| < \eta_2/2$ for every $\beta \in P''$ when $0 < \epsilon \leq \epsilon_1$ (see (3.23)).

Further, there is a number $t_\epsilon' \in (t_\epsilon, t_1]$ such that $\int_{t_\epsilon}^{t_\epsilon'} m_0(t) dt < \eta_2/4$.

But this means (see (3.13)) that the solutions $x(t; \beta, \epsilon)$ of (3.16), for every $\beta \in P''$, can be extended beyond t_ϵ to t_ϵ' in such a way that, for every $t \in [t_1, t_\epsilon']$, $x(t; \beta, \epsilon) \in X_1$ and $|x(t; \beta, \epsilon) - z(t)| < \eta_0/2$, contradicting the definition of t_ϵ .

Thus, we have shown that, if $0 \leq \epsilon \leq \epsilon_1$, then $I_\epsilon = I$; i.e., $x(t; \beta, \epsilon)$, for every $\beta \in P''$, is defined on all of I and takes on values in X_1 , and $|x(t; \beta, \epsilon) - z(t)| < \eta_0/2$ for all $t \in I$. Also, (3.20) and (3.23) are satisfied for all $(t, \beta) \in I \times P''$ whenever $\epsilon \in [0, \epsilon_1]$, which immediately implies that (3.19) holds.

Let us now proceed to verify (3.17). We shall henceforth suppose that $0 < \epsilon \leq \epsilon_1$. Thus, for all $(t, \beta) \in I \times P^\nu$, $|x(t; \beta, \epsilon) - z(t)| < \eta_0/2$, so that

$$(3.24) \quad \begin{aligned} |x(t; \beta, \epsilon) - x(t; \beta', \epsilon)| &\leq |x(t; \beta, \epsilon) - z(t)| + |x(t; \beta', \epsilon) - z(t)| \\ &< \eta_0 \quad \text{for all } (t, \beta, \beta') \in I \times P^\nu \times P^\nu, \end{aligned}$$

and the entire line segment joining $x(t; \beta, \epsilon)$ with $z(t)$ lies in $X_0 \subset G$. Since $g(x, t; \beta, \epsilon) \in \Gamma$, and the elements of Γ are of class C^1 with respect to x in G , it now follows that

$$g^*(x(t; \beta, \epsilon), t) - g^*(z(t), t) = \tilde{g}_x(t; \beta, \epsilon)[x(t; \beta, \epsilon) - z(t)]$$

for all $(t, \beta) \in I \times P^\nu$,

where $\tilde{g}_x(t; \beta, \epsilon)$ is an $n \times n$ matrix-valued function which differs from $g_x^*(z(t), t)$ only in that the partial derivatives are evaluated at some point on the line segment joining $z(t)$ and $x(t; \beta, \epsilon)$ rather than at $z(t)$. Consequently (see (3.21) and (3.1)),

$$(3.25) \quad \begin{aligned} \frac{x(t; \beta, \epsilon) - z(t)}{\epsilon} &= \sum_{i=1}^p \beta_i \xi_i + \int_{t_1}^t \delta g(z(s), s; \beta) ds \\ &+ \int_{t_1}^t g_x^*(z(s), s) \frac{x(s; \beta, \epsilon) - z(s)}{\epsilon} ds + \lambda(t; \beta, \epsilon) \end{aligned}$$

for all $(t, \beta) \in I \times P^\nu$, where

$$\begin{aligned} \lambda(t; \beta, \epsilon) &= \int_{t_1}^t [\delta g(x(s; \beta, \epsilon), s; \beta) - \delta g(z(s), s; \beta)] ds \\ &+ \int_{t_1}^t [\tilde{g}_x(s; \beta, \epsilon) - g_x^*(z(s), s)] \frac{x(s; \beta, \epsilon) - z(s)}{\epsilon} ds \\ &+ \frac{1}{\epsilon} \int_{t_1}^t \delta^2 g(x(s; \beta, \epsilon), s; \beta, \epsilon) ds. \end{aligned}$$

Let us show that

$$(3.26) \quad \lambda(t; \beta, \epsilon) \xrightarrow{\epsilon \rightarrow 0^+} 0 \quad \text{uniformly with respect to } (t, \beta) \in I \times P^\nu.$$

Now, by virtue of (3.20) and (3.23), we have that

$$(3.27) \quad \begin{aligned} |\lambda(t; \beta, \epsilon)| &\leq \int_I |\delta g(x(s; \beta, \epsilon), s; \beta) - \delta g(z(s), s; \beta)| ds \\ &+ \tilde{\eta} \int_I |\tilde{g}_x(s; \beta, \epsilon) - g_x^*(z(s), s)| ds + \epsilon \end{aligned}$$

for every $(t, \beta) \in I \times P^\nu$. Further, for each fixed $s \in I$, the functions $\delta g(x, s; \beta)$ for $\beta \in P^\nu$ and $g_x^*(x, s)$ are uniformly equicontinuous with respect

to $x \in X_1$. It now follows from (3.19) and (3.22) that there exist measurable, real-valued functions $\zeta(s; \epsilon)$ defined on I for each ϵ such that (1) the integrands in the right-hand side of (3.27) are dominated by $\zeta(s; \epsilon)$ for every s, β , and ϵ , (2) $\zeta(s; \epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$ for each $s \in I$, and (3) $\zeta(s; \epsilon) \leq 4m_0(s)$ for each s and ϵ . Applying the Lebesgue dominated convergence theorem, we conclude that (3.26) holds. Let $\hat{\lambda}(\epsilon) = \max_{(t, \beta) \in I \times P^v} |\lambda(t; \beta, \epsilon)|$, so that

$$(3.28) \quad \hat{\lambda}(\epsilon) \xrightarrow{\epsilon \rightarrow 0^+} 0.$$

If we set

$$(3.29) \quad q(t; \beta, \epsilon) = \frac{x(t; \beta, \epsilon) - z(t)}{\epsilon} - \delta x(t; \beta),$$

it follows from (3.25) and (3.7) that, for all $(t, \beta) \in I \times P^v$,

$$q(t; \beta, \epsilon) = \int_{t_1}^t g_x^*(z(s), s) q(s; \beta, \epsilon) ds + \lambda(t; \beta, \epsilon).$$

Hence (see (3.22)),

$$|q(t; \beta, \epsilon)| \leq \int_{t_1}^t m_0(s) |q(s; \beta, \epsilon)| ds + \hat{\lambda}(\epsilon),$$

and, by virtue of Gronwall's inequality, we conclude that

$$(3.30) \quad |q(t; \beta, \epsilon)| \leq \hat{\lambda}(\epsilon) \exp \left(\int_I m_0(s) ds \right) \text{ for all } (t, \beta) \in I \times P^v.$$

Recalling that $\delta x(t; \beta) = \sum_{i=1}^v \beta_i \delta x_{\xi_i, h_i}(t)$, (3.17) now follows from (3.28)–(3.30).

We now turn to the verification of (3.18). It is a consequence of (3.16) that

$$\begin{aligned} x(t; \beta', \epsilon) - x(t; \beta, \epsilon) &= \epsilon \sum_{i=1}^v (\beta_i' - \beta_i) \xi_i \\ &\quad + \int_{t_1}^t [g(x(s; \beta', \epsilon), s; \beta', \epsilon) - g(x(s; \beta, \epsilon), s; \beta, \epsilon)] ds \end{aligned}$$

for all $(t, \beta, \beta') \in I \times P^v \times P^v$. Since ϵ is fixed in the remainder of the proof, we shall, for ease of notation, drop it as an argument for the functions x and g . Then, clearly, for all $(t, \beta, \beta') \in I \times P^v \times P^v$,

$$\begin{aligned} |x(t; \beta') - x(t; \beta)| &\leq \epsilon \sum_{i=1}^v |\beta_i' - \beta_i| |\xi_i| \\ (3.31) \quad &+ \left| \int_{t_1}^t [g(x(s; \beta'), s; \beta') - g(x(s; \beta), s; \beta)] ds \right| \end{aligned}$$

$$+ \int_{t_1}^t |g(x(s; \beta'), s; \beta) - g(x(s; \beta), s; \beta)| ds.$$

It follows from (3.24) that, for all $(s, \beta, \beta') \in I \times P^v \times P^v$, the entire line segment joining $x(s; \beta)$ with $x(s; \beta')$ lies in X_0 . Consequently (see (3.13)),

$$(3.32) \quad \begin{aligned} & |g(x(s; \beta'), s; \beta) - g(x(s; \beta), s; \beta)| \\ & \leq \sup_{x \in X_0} |g_x(x, s; \beta)| |x(s; \beta') - x(s; \beta)| \\ & \leq m_0(s) |x(s; \beta') - x(s; \beta)|. \end{aligned}$$

Further, it follows from (3.15) and Lemma 3.1 that

$$(3.33) \quad \left| \int_{t_1}^t [g(x(s; \beta'), s; \beta') - g(x(s; \beta'), s; \beta)] ds \right| \xrightarrow[\beta' \in P^v]{\beta' \rightarrow \beta} 0$$

uniformly with respect to $t \in I$.

Using relations (3.31)–(3.33), and applying the Gronwall inequality in the same way as before, we arrive at the desired relation (3.18). This completes the proof of Theorem 3.1.

Note 3.1. An important class of quasiconvex functions may be defined as follows. Let $f(x, u, t)$ be a function from $G \times U \times I$ into R^n (recall that G is an open set in R^n), where U is an arbitrary fixed set in R^r . Let us suppose that f is of class C^1 with respect to $x \in G$ and measurable with respect to $(u, t) \in U \times I$ (in the sense that the preimage of every Borel set is a Borel set) for each fixed $x \in G$. Let U_t , for each $t \in I$, be a fixed subset of U , and let Ω^* denote the set of all measurable, essentially bounded functions $u(t)$ from I into R^r such that $u(t) \in U_t$ for almost all $t \in I$. Finally, suppose that for every compact subset X of G and each $\bar{u} \in \Omega^*$, there exists a function $\bar{m}(t) \in L^1(I)$ such that $|f(x, \bar{u}(t), t)| + |f_x(x, \bar{u}(t), t)| \leq \bar{m}(t)$ for all $(t, x) \in I \times X$. For each $u \in \Omega^*$, denote the function $f(x, u(t), t)$ from $G \times I$ to R^n by $g^u(x, t)$. It was proved in [3, §4] that the set $\{g^u : u \in \Omega^*\} = \Gamma^*$ is quasiconvex. (Recall that if a class of functions is quasiconvex in the sense of [3] then it is also quasiconvex in the sense of our Definition 3.2.) Indeed, it is this class that motivated the definition of quasiconvexity.

4. A first-order, convex approximation in a finite-dimensional space.

In this section we shall construct a first-order convex approximation to a set in a finite-dimensional space which corresponds to the solutions of a class of difference equations.

Thus, let us consider the space R^{nk} , where n and k are preassigned positive integers, under the conventional Euclidean norm topology. We shall, for convenience, represent the elements $x \in R^{nk}$ as k -tuples of vectors in

R^n , i.e., $x = (\xi_1, \dots, \xi_k)$, where $\xi_i \in R^n$ for $i = 1, \dots, k$. Let $R_0^{nk} = \{x = (\xi_1, \dots, \xi_k) : \xi_i \in R^n \text{ for } i = 1, \dots, k, \xi_1 = 0\}$.

Let G be an open subset of R^n , let I_0 denote the set of integers $\{1, \dots, k-1\}$, let \mathcal{G}_0 denote the set of all functions $g(\xi, i)$ from $G \times I_0$ into R^n which are of class C^1 with respect to $\xi \in G$, and let Γ be a fixed convex subset of \mathcal{G}_0 (i.e., if $g_1 \in \Gamma$ and $g_2 \in \Gamma$, then $(\alpha g_1 + \beta g_2) \in \Gamma$ whenever $(\alpha, \beta) \in P^2$).

Now let \tilde{Q}' denote the set of all $x = (\xi_1, \dots, \xi_k) \in R^{nk}$ such that $\xi_i \in G$ for each i , and such that, for some $g \in \Gamma$,

$$\xi_{i+1} = g(\xi_i, i) \quad \text{for each } i = 1, \dots, k-1.$$

Let $z \in \tilde{Q}'$, so that $z = (\zeta_1, \dots, \zeta_k)$ with $\zeta_i \in G$ for each i and

$$(4.1) \quad \zeta_{i+1} = g^*(\zeta_i, i) \quad \text{for } i = 1, \dots, k-1, \quad \text{where } g^* \in \Gamma.$$

Also, let $\tilde{Q} = \tilde{Q}' - z$, and let \tilde{K} denote the set of all k -tuples $(\xi_1, \dots, \xi_k) \in R^{nk}$ such that, for some $g \in \Gamma$,

$$(4.2) \quad \xi_i = \Phi_{i,1}\xi_1 + \sum_{j=1}^{i-1} \Phi_{i,j+1}[g(\zeta_j, j) - g^*(\zeta_j, j)] \quad \text{for } i = 2, \dots, k,$$

where the $\Phi_{i,j}$, for $1 \leq j \leq i \leq k$, are the $n \times n$ matrices defined by the recurrence relations

$$(4.3) \quad \Phi_{i,j} = \Phi_{i,j+1} \frac{\partial g^*(\zeta_j, j)}{\partial \xi} \quad \text{for } 1 \leq j \leq i-1 \leq k-1,$$

$$\Phi_{i,i} = \text{the identity for } i = 1, \dots, k.$$

Then it can be shown (the arguments are similar to, but much simpler than, those used to prove Theorem 3.1) that \tilde{K} is a first-order, convex approximation to \tilde{Q} in R^{nk} , and that $\tilde{K} \cap R_0^{nk}$ is a first-order convex approximation to $(\tilde{Q} \cap R_0^{nk})$ in R_0^{nk} .

Note 4.1. A convex subset of \mathcal{G}_0 , which is important in applications, may be defined as follows. Let $f(\xi, u, i)$ be a function from $G \times U \times I_0$ into R^n , where G and I_0 are defined as before, and U is an arbitrary fixed subset of R^n . Let us suppose that f is of class C^1 with respect to $\xi \in G$, and let U_1, \dots, U_{k-1} be fixed subsets of U such that $f(\xi, U_i, i) = \{f(\xi, v, i) : v \in U_i\}$ is a convex subset of R^n for every $\xi \in G$ and $i \in I_0$. Let $\tilde{\Omega}$ denote the set of all functions $u(i)$ from I_0 into U such that $u(i) \in U_i$ for every $i \in I_0$. For each function $u \in \tilde{\Omega}$, let $g^u(\xi, i)$ denote the function from $G \times I_0$ into R^n defined by the relation

$$g^u(\xi, i) = f(\xi, u(i), i).$$

Then it is easily seen that the set $\tilde{\Gamma} = \{g^u : u \in \tilde{\Omega}\}$ is a convex subset of \mathcal{G}_0 .

5. Some particular functionals and their differentials. In this section we shall describe some functionals φ , defined on regions in certain locally convex, linear topological spaces, which possess differentials in the sense that (5.1) (or (5.8)) holds, with l a linear (or c a convex) continuous functional defined on the underlying space.

We first note that, if the underlying space \mathfrak{J} is R^n , and if φ is a functional defined on a subset \tilde{G} of R^n such that φ possesses a differential l (in the ordinary sense, so that l is, in particular, linear and continuous) at a point $z \in (\text{interior of } \tilde{G})$, then it is easily seen that

$$(5.1) \quad \frac{\varphi(z + \epsilon y) - \varphi(z)}{\epsilon} \xrightarrow[\epsilon \rightarrow 0]{y \rightarrow x} l(x) \quad \text{for all } x \in \mathfrak{J}.$$

As a second example, consider the case where the underlying space \mathfrak{J} is \mathcal{S} , \mathcal{S}_0 , or \mathcal{C} (as defined in §2). Let τ_1, \dots, τ_k be fixed distinct points in $I = [t_1, t_2]$, let G_0 be a nonempty open set in R^{nk} , and let $\chi(\xi_1, \dots, \xi_k)$ be a function from G_0 into R^1 (each ξ_j denotes an n -vector) of class C^1 . If $W_0 = \{x: x \in \mathfrak{J}, (x(\tau_1), \dots, x(\tau_k)) \in G_0\}$, then it is clear that W_0 is a nonempty open set in \mathfrak{J} (whether \mathfrak{J} is \mathcal{C} , \mathcal{S} , or \mathcal{S}_0). Now define the function φ from W_0 into R^1 as follows:

$$(5.2) \quad \varphi(x) = \chi(x(\tau_1), \dots, x(\tau_k)).$$

It is clear that φ is continuous on W_0 and that, if $z \in W_0$, then (5.1) is satisfied, where

$$(5.3) \quad l(x) = \sum_{j=1}^k \frac{\partial \chi(z(\tau_1), \dots, z(\tau_k))}{\partial \xi_j} \cdot x(\tau_j), \quad l \in \mathfrak{J}^*.$$

(The preceding statement holds whether \mathfrak{J} is \mathcal{C} , \mathcal{S} , or \mathcal{S}_0 .)

In the third example we shall consider, the underlying space \mathfrak{J} is $\mathcal{C} \times R^k$, where \mathcal{C} is defined as in §2 and k is a fixed positive integer. Let \tilde{G} be a nonempty open set in $R^{nk} \times I^k$, and let $\theta(\xi_1, \dots, \xi_k, \sigma_1, \dots, \sigma_k)$ be a function from \tilde{G} into R^1 of class C^1 (each ξ_j denotes an n -vector and each σ_j a real number). If we set

$$(5.4) \quad W_1 = \{x = (x, \tau_1, \dots, \tau_k): x \in \mathcal{C}, (x(\tilde{\tau}_1), \dots, x(\tilde{\tau}_k), \tau_1, \dots, \tau_k) \in \tilde{G}\},$$

where

$$(5.5) \quad \tilde{\tau}_i = \tau_i \quad \text{if } \tau_i \in I, \quad \tilde{\tau}_i = t_1 \quad \text{if } \tau_i < t_1, \quad \tilde{\tau}_i = t_2 \quad \text{if } \tau_i > t_2, \\ i = 1, \dots, k,$$

then W_1 is easily seen to be an open set in \mathfrak{J} . Now define the function φ from W_1 into R^1 as follows:

$$(5.6) \quad \varphi(\mathbf{x}) = \varphi(x, \tau_1, \dots, \tau_k) = \theta(x(\tilde{\tau}_1), \dots, x(\tilde{\tau}_k), \tau_1, \dots, \tau_k),$$

where the $\tilde{\tau}_i$ are defined by (5.5).

It is easy to show that φ is continuous on W_1 , and that, if $\mathbf{z} = (z, \tau_1^*, \dots, \tau_k^*) \in W_1$, where $t_1 < \tau_i^* < t_2$ and $z(t)$ is differentiable at τ_i^* for each $i = 1, \dots, k$, then (5.1) holds (with z, y , and x replaced by \mathbf{z}, y , and \mathbf{x} , respectively), where

$$(5.7) \quad \begin{aligned} l(\mathbf{x}) &= l(x, \tau_1, \dots, \tau_k) \\ &= \sum_{j=1}^k \frac{\partial \theta(z(\tau_1^*), \dots, z(\tau_k^*), \tau_1^*, \dots, \tau_k^*)}{\partial z_j} \cdot [x(\tau_j^*) + \tau_j \dot{z}(\tau_j^*)] \\ &\quad + \sum_{j=1}^k \tau_j \frac{\partial \theta(z(\tau_1^*), \dots, z(\tau_k^*), \tau_1^*, \dots, \tau_k^*)}{\partial \sigma_j}, \quad l \in \mathfrak{J}^*. \end{aligned}$$

We point out that the functional φ defined in the preceding example does not in general possess a Fréchet differential.

For our last example, we shall consider a functional φ which does not generally possess a linear continuous differential l in the sense that (5.1) holds. Instead we shall show that for each z in the domain of φ there is a convex, continuous functional c , defined on the entire underlying space \mathfrak{J} , such that

$$(5.8) \quad \frac{\varphi(z + \epsilon y) - \varphi(z)}{\epsilon} \xrightarrow[\substack{\epsilon \rightarrow 0^+ \\ y \rightarrow x}}{c(x)} \quad \text{for every } x \in \mathfrak{J}.$$

In this example, the underlying space \mathfrak{J} is \mathcal{C} . Let G be a nonempty open set in R^n and let $\tilde{g}(x, t)$ be a continuous function from $G \times I$ into R^1 such that $\tilde{g}_x(x, t)$ is defined and continuous on $G \times I$. Let $\mathcal{C}_G = \{x : x \in \mathcal{C}, x(t) \in G \text{ for all } t \in I\}$. Evidently, \mathcal{C}_G is a nonempty open set in \mathcal{C} . Let \tilde{I} be a fixed subset of I . Then define the functional φ from \mathcal{C}_G into R^1 as follows:

$$(5.9) \quad \varphi(x) = \sup_{t \in \tilde{I}} \tilde{g}(x(t), t).$$

It is easily verified that φ is continuous on \mathcal{C}_G .

Let $z \in \mathcal{C}_G$. We shall show that (5.8) holds with

$$(5.10) \quad c(x) = \sup_{t \in I_e} [\tilde{g}_x(z(t), t) \cdot x(t)],$$

where

$$(5.11) \quad I_e = \{t : t \in I, \tilde{g}(z(t), t) = \varphi(z)\} \cap (\text{closure of } \tilde{I}).$$

For ease of notation, and without loss of generality, we shall suppose that

$$(5.12) \quad \varphi(z) = \sup_{t \in \tilde{I}} \tilde{g}(z(t), t) = 0.$$

It is obvious that I_e is closed and not empty, and that c is a convex, continuous (but generally nonlinear) functional defined on \mathcal{C} . Further, if $\tilde{g}_x(z(t), t) \neq 0$ for all $t \in I_e$, there is at least one $x \in \mathcal{C}$, $x \neq 0$, such that $c(x) < 0$ (namely, let $x(t) = -\tilde{g}_x(z(t), t)$).

Let us verify (5.8). Let $x \in \mathcal{C}$ be fixed but arbitrary. Then the following equation holds whenever $y \in \mathcal{C}$, $\epsilon > 0$, and $\|y - x\|$ are sufficiently small:

$$(5.13) \quad \begin{aligned} \varphi(z + \epsilon y) &= \sup_{t \in \tilde{I}} \tilde{g}(z(t) + \epsilon y(t), t) \\ &= \sup_{t \in \tilde{I}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot y(t)] \\ &= \sup_{t \in \tilde{I}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t) + \epsilon \zeta(t; y, \epsilon)], \end{aligned}$$

where

$$\begin{aligned} \zeta(t; y, \epsilon) &= \tilde{g}_x(z(t), t) \cdot [y(t) - x(t)] \\ &\quad + [\tilde{g}_x(z(t) + \epsilon \theta_{\epsilon, t} y(t), t) - \tilde{g}_x(z(t), t)] \cdot y(t), \end{aligned}$$

and $\theta_{\epsilon, t}$ is a number (which depends on y as well as ϵ and t) such that $0 \leq \theta_{\epsilon, t} \leq 1$. It follows from our hypotheses on the function \tilde{g} that

$$(5.14) \quad \sup_{t \in I} |\zeta(t; y, \epsilon)| \xrightarrow[\substack{\epsilon \rightarrow 0^+ \\ y \rightarrow x}]{} 0.$$

Further (see (5.10)–(5.13)),

$$\begin{aligned} \varphi(z + \epsilon y) &\geq \sup_{t \in I_e} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t) + \epsilon \zeta(t; y, \epsilon)] \\ &\geq \epsilon \sup_{t \in I_e} [\tilde{g}_x(z(t), t) \cdot x(t) + \zeta(t; y, \epsilon)] \\ &\geq \epsilon [c(x) - \sup_{t \in I} |\zeta(t; y, \epsilon)|], \end{aligned}$$

from which it follows, by virtue of (5.14), that

$$(5.15) \quad \liminf_{\substack{\epsilon \rightarrow 0^+ \\ y \rightarrow x}} \frac{\varphi(z + \epsilon y)}{\epsilon} \geq c(x).$$

For each $\delta > 0$, let $N_\delta = \{t + \tau : t \in I_e, |\tau| < \delta\} \cap \tilde{I}$. Let $\eta > 0$ be arbitrary but fixed. Then there is a $\delta_1 > 0$ such that

$$(5.16) \quad \tilde{g}_x(z(t), t) \cdot x(t) \leq c(x) + \eta \quad \text{whenever } t \in N_{\delta_1}$$

(see (5.10)). Let us show that if $\epsilon > 0$ is sufficiently small then

$$(5.17) \quad \sup_{t \in \tilde{I}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t)] \leq \sup_{t \in N_{\delta_1}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t)].$$

If $\tilde{I} \subset N_{\delta_1}$, then (5.17) is immediately seen to hold. Thus, let us suppose that $\tilde{I} = \tilde{I} \cap (\text{complement of } N_{\delta_1})$ is not empty, and let $\delta_2 = \sup_{t \in \tilde{I}} \tilde{g}(z(t), t)$. Let $\bar{t} \in (\text{closure of } \tilde{I})$ be such that $\tilde{g}(z(\bar{t}), \bar{t}) = \delta_2$; clearly, $\bar{t} \in (\text{closure of } \tilde{I})$. It then follows from the definitions of N_{δ_1} and \tilde{I} that $\bar{t} \notin I_e$, so that (see (5.11) and (5.12)) $\delta_2 < 0$. Consequently, since $I_e \subset (\text{closure of } N_{\delta_1})$,

$$\sup_{t \in N_{\delta_1}} \tilde{g}(z(t), t) \geq \sup_{t \in I_e} \tilde{g}(z(t), t) = 0 > \sup_{t \in \tilde{I}} \tilde{g}(z(t), t) = \delta_2.$$

Hence, if $\epsilon > 0$ is sufficiently small,

$$\sup_{\substack{t \in \tilde{I} \\ t \notin N_{\delta_1}}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t)] \leq \sup_{t \in N_{\delta_1}} [\tilde{g}(z(t), t) + \epsilon \tilde{g}_x(z(t), t) \cdot x(t)],$$

which immediately implies (5.17). Consequently, if $\epsilon > 0$ and $\|y - x\|$ are sufficiently small (where $y \in \mathcal{C}$), it follows from (5.13), (5.17), (5.16) and (5.12) that

$$\varphi(z + \epsilon y) \leq \epsilon[c(x) + \eta + \sup_{t \in I} |\dot{z}(t; y, \epsilon)|],$$

which, together with (5.14), imply that

$$(5.18) \quad \limsup_{\substack{\epsilon \rightarrow 0^+ \\ y \rightarrow x}} \frac{\varphi(z + \epsilon y)}{\epsilon} \leq c(x) + \eta.$$

Since $\eta > 0$ is arbitrary, (5.8) now follows directly from (5.15), (5.18) and (5.12).

6. Applications to optimal control problems. In this section we shall consider some optimal control problems which fall under the category of the canonical optimization problem described in [1, §4]. Under suitable hypotheses, we shall show that the solutions of this problem are smoothly regular, and, appealing to Theorems 4.2 and 4.5 in [1], and making use of the results of §§2, 4 and 5 of the present article, we shall obtain particular necessary conditions for the solutions of our optimal control problems.

Let the sets I , G , U , Ω , and Q_0' , the function f , and the linear topological spaces \mathcal{C} , \mathcal{S} , and \mathcal{S}_0 be defined as in §2 and satisfy the hypotheses described

therein. Then consider the following conventional optimal control problem.

Problem 6.1. Let $\mathfrak{J} = \mathfrak{S}_0$, and let $\varphi_0, \varphi_1, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (5.2) where $k = 1$, $\tau_1 = t_2$, and the corresponding real-valued functions χ_i have a common open domain $G_0 \subset G$ and are of class C^1 . Then find an element $x \in W \cap Q_0'$ (where $W \subset \mathfrak{S}_0$ is the common domain of the φ_i) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$ and such that $\varphi_0(x)$ is minimized.

It is clear that Problem 6.1 falls under the category of the canonical optimization problem described in [1, §4] (with $\mu = 0$ and $Q' = Q_0' \cap \mathfrak{S}_0$).

Let z be a solution of this problem, so that z satisfies (2.3) for some $u^* \in \Omega$. Let

$$(6.1) \quad \chi_i' = \frac{\partial \chi_i(z(t_2))}{\partial \xi}, \quad i = 0, 1, \dots, m,$$

and let us suppose that the vectors $\chi_0', \chi_1', \dots, \chi_m'$ are linearly independent.

It is now easy to verify on the basis of the results of §5 that z is a smoothly regular solution of our canonical optimization problem (in the sense of [1, Definition 4.4]). Inasmuch as K_0 , given by (2.8), is a first-order, convex approximation to $(Q_0' - z) \cap \mathfrak{S}_0$ in \mathfrak{S}_0 (this was shown in §2), we can appeal to Theorem 4.2 in [1] and conclude that there exist real numbers α_i , $i = 0, 1, \dots, m$, not all zero, such that

$$(6.2) \quad \sum_{i=0}^m \alpha_i l_i(x) \leq 0 \quad \text{for all } x \in K_0, \quad \alpha_0 \leq 0,$$

where (see (5.3) and (6.1)) $l_i(x) = \chi_i' \cdot x(t_2)$ for $i = 0, 1, \dots, m$.

Since $x_{s,v,0} \in K_0$ whenever $(s, v) \in I_1 \times U$ (see (2.8)), (6.2) implies that $\sum_{i=0}^m \alpha_i \chi_i' \cdot x_{s,v,0}(t_2) \leq 0$ for all $(s, v) \in I_1 \times U$. Further, it follows from (2.5)–(2.7) that, if $t_1 \leq s < t_2$, then

$$(6.3) \quad x_{s,v,0}(t_2) = \Phi(t_2)\Phi^{-1}(s)[f(z(s), v) - f(z(s), u^*(s))],$$

where $\Phi(t)$ is any nonsingular, absolutely continuous $n \times n$ matrix function which satisfies the equation

$$(6.4) \quad \dot{\Phi}(t) = f_x(z(t), u^*(t))\Phi(t) \quad \text{for almost all } t \in I.$$

Consequently, if we set $\psi(t) = \sum_{i=0}^m \alpha_i \chi_i' \Phi(t_2)\Phi^{-1}(t)$ for $t \in I$, and consider both ψ and the χ_i' to be row vectors, we can conclude that ψ is an absolutely continuous function and that (recall that $\text{meas } I_1 = \text{meas } I$)

$$(6.5) \quad \psi(t)f(z(t), u^*(t)) = \max_{v \in U} \psi(t)f(z(t), v) \quad \text{for almost all } t \in I,$$

$$(6.6) \quad \dot{\psi}(t) = -\psi(t)f_x(z(t), u^*(t)) \quad \text{for almost all } t \in I,$$

$$(6.7) \quad \psi(t_2) = \sum_{i=0}^m \alpha_i \chi_i', \quad \alpha_0 \leq 0.$$

Since the α_i do not all vanish and the χ_i' , $i = 0, \dots, m$, are linearly independent by hypothesis, it follows from (6.7) that $\psi(t) \neq 0$.

Relations (6.5) and (6.6) are immediately recognizable as the Pontryagin maximum principle [2], and (6.7) (with (6.1)) is the associated transversality condition. (In [2], it was assumed that the functions $\chi_i(\xi)$ for $i > 0$ as well as $f(\xi, u)$ are independent of ξ^1 , and that $\chi_0(\xi) = \xi^1$. With these additional hypotheses, (6.6) and (6.7) imply that $\psi^1(t) \equiv \alpha_0 \leq 0$, as was stated in [2].)

Thus we have shown that if $z \in S_0$ is a solution of Problem 6.1 satisfying (2.3) for some $u^* \in \Omega$, such that the vectors χ_i' , $i = 0, 1, \dots, m$, defined by (6.1) are linearly independent, then there exists an absolutely continuous, nonzero, (row) n -vector valued function $\psi(t)$ defined on I such that (6.5)–(6.7) hold for some constants $\alpha_0, \alpha_1, \dots, \alpha_m$.

Note 6.1. Problem 6.1 is a conventional fixed-time optimal control problem with fixed left-hand endpoint. Admittedly, the left-hand initial condition has the particular form $x(t_1) = 0$ (recall that we are seeking solutions in S_0), but this restriction is only apparent inasmuch as a problem where the initial condition is of the form $x(t_1) = \xi_0$, with fixed ξ_0 , can be transformed into a zero initial condition problem by a simple change of variables. In order to consider problems with a variable left-hand endpoint, Problem 6.1 must be modified by replacing S_0 by S , and by considering that the functionals φ_i have the form of (5.2), where $k = 2$, $\tau_1 = t_1$, and $\tau_2 = t_2$, and the corresponding functions χ_i are of class C^1 and have domain $G \times G$. Then, reasoning almost exactly as in Problem 6.1, but with K_0 replaced by K_1 , one obtains necessary conditions for solutions of this problem which are as follows: (6.5) and (6.6) remain unchanged; (6.7) is essentially the same and supplemented by an analogous transversality condition at the left-hand endpoint.

Note 6.2. Problem 6.1 may be further generalized as follows: Again replace S_0 by S . In addition to the functionals φ_i , $i = 0, 1, \dots, m$, functionals φ_{-i} , $i = 1, \dots, \mu$, are given. Further, the φ_i , $i = -\mu, \dots, 0, \dots, m$, are of the form of (5.2), but with k not necessarily equal to 1 and with the τ_i , $i = 1, \dots, k$, arbitrary, fixed points in I . However, we still assume that the corresponding functions χ_i have a common domain $G_0 \subset G^k$ and are of class C^1 . Then the problem consists in finding an element $x \in W \cap Q_0'$ (where $W \subset S$ is the common domain of the φ_i) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$, $\varphi_{-i}(x) \leq 0$ for $i = 1, \dots, \mu$, and such that $\varphi_0(x)$ is minimized. If z is a solution of this problem such that the Jacobian of the χ_i at $(z(\tau_1), \dots, z(\tau_k))$ satisfies a certain regularity condition, it is possible to apply Theorem 4.2 of [1] and obtain necessary conditions analogous

to (6.5)–(6.7). If we return to the space S_0 and if $k = 1$ and $\tau_1 = t_2$, these necessary conditions coincide with those obtained for Problem 6.1, except that (6.7) must be replaced by the conditions

$$(6.8) \quad \begin{aligned} \psi(t_2) &= \sum_{i=-\mu}^m \alpha_i \chi_i', & \alpha_{-i} &\leq 0 \quad \text{for } i = 0, 1, \dots, \mu, \\ \alpha_{-i} &= 0 \quad \text{if } i > 1 \quad \text{and} \quad \chi_{-i}(z(t_2)) < 0, \end{aligned}$$

and (6.1) holds for every $i = -\mu, \dots, m$.

In §7 we shall consider a general optimization problem from the viewpoint of quasiconvexity which will include, as special cases, Problem 6.1 together with the generalizations described in Notes 6.1 and 6.2, as well as optimal control problems with “free” initial and terminal times.

Note 6.3. If, in the statement of Problem 6.1, the roles of φ_0 and one of the φ_j with $j > 0$ are interchanged, then the form of the necessary conditions for solutions of this problem is essentially unchanged. (However, the condition $\alpha_0 \leq 0$ must be replaced by $\alpha_j \leq 0$.) This invariance was pointed out in [1, Note 4.4].

Note 6.4. If, in Problem 6.1, $m = 0$ (i.e., if the constraints $\varphi_i(x) = 0$ for $i = 1, \dots, m$ are omitted), we obtain what in [1] was referred to as a simple optimization problem, and which is commonly known as a problem with a “free right-hand endpoint.” Appealing to Theorem 4.5 of [1], we can easily derive necessary conditions satisfied by solutions of this problem. These differ from those satisfied by solutions of Problem 6.1 only in that (6.7) must be replaced by the conditions $\psi(t_2) = \alpha_0 \chi_0', \alpha_0 \leq 0$.

Let us now return to the original Problem 6.1, and let us make the additional hypothesis that $f(x, u)$ is of class C^1 with respect to $(x, u) \in G \times U$. We shall also suppose \mathfrak{J} is $\mathfrak{C} \cap S_0$ rather than S_0 .

Let z be a solution of this problem, so that (2.3) holds for some $u^* \in \Omega$. We shall again suppose that the vectors $\chi_i', i = 0, \dots, m$, defined by (6.1) are linearly independent, and, in addition, that $u^*(t) \in (\text{interior of } U)$ for all $t \in I$. (The last hypothesis will be automatically satisfied if U is open.) It was pointed out in §2 that, under our hypotheses, $K_0' = \{y_u : y_u \text{ a solution of (2.9) with } u \in L_r^\infty(I)\}$ is a first-order, convex approximation to $(Q_0' - z) \cap S_0$. Once again appealing to Theorem 4.2 of [1], we can conclude that there exist real numbers $\alpha_i, i = 0, 1, \dots, m$, not all zero, such that

$$(6.9) \quad \sum_{i=0}^m \alpha_i l_i(y_u) \leq 0 \quad \text{for all } y_u \in K_0', \quad \alpha_0 \leq 0,$$

where $l_i(y_u) = \chi_i' \cdot y_u(t_2)$ for $i = 0, \dots, m$. It follows from (2.9) that

$$y_u(t_2) = \Phi(t_2) \int_{t_1}^{t_2} \Phi^{-1}(s) f_u(z(s), u^*(s)) u(s) \, ds,$$

where Φ is any nonsingular absolutely continuous $n \times n$ matrix function which satisfies (6.4). Thus, (6.9) can be rewritten in the form

$$\sum_{i=0}^m \alpha_i \chi_i' \cdot \Phi(t_2) \int_{t_1}^{t_2} \Phi^{-1}(s) f_u(z(s), u^*(s)) u(s) ds \leq 0$$

for all $u \in L_r^\infty(I)$,

or, if we set $\psi(s) = \sum_{i=0}^m \alpha_i \chi_i' \Phi(t_2) \Phi^{-1}(s)$ for all $s \in I$ (and consider ψ and the χ_i' to be row vectors),

$$(6.10) \quad \int_{t_1}^{t_2} \psi(s) f_u(z(s), u^*(s)) u(s) ds \leq 0 \quad \text{for all } u \in L_r^\infty(I).$$

But (6.10) is possible only if $\psi(t) f_u(z(t), u^*(t)) = 0$ whenever $t \in I$, i.e.,

$$(6.11) \quad \psi(t) f_u(z(t), u^*(t)) = 0 \quad \text{for almost all } t \in I.$$

Finally, (6.6), (6.7), and the fact that $\psi \neq 0$ follow as before.

Thus, if z is a solution of Problem 6.1 such that all of the above indicated hypotheses are satisfied, there exists a nonzero, absolutely continuous, (row) n -vector valued function $\psi(t)$ defined on I such that (6.6), (6.7) and (6.11) are satisfied for some constants $\alpha_0, \dots, \alpha_m$.

It is easily seen that (6.5) implies (6.11) if f and u^* satisfy the stronger hypotheses described above. The only reason we have presented a separate derivation of the weaker necessary condition is to illustrate the generality of our method of obtaining necessary conditions.

We shall now consider two problems which arise in the optimal control of sampled-data control systems.

Let the sets $I, G, U, Q_1', Q_2', I_0, \Omega_1$, and Ω_2 and the function f be defined as in §2 and satisfy the hypotheses (including the assumption that f is of class C^1 with respect to both x and u) described therein.

Problem 6.2. Let $\mathfrak{J} = S_0 \cap \mathfrak{C}$ and let $\varphi_0, \varphi_1, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (5.2), with $k = 1$ and $\tau_1 = t_2$, where the corresponding real-valued functions χ_i have a common open domain $G_0 \subset G$, and are of class C^1 . Then find an element $x \in W \cap Q_1'$ (where $W \subset S_0$ is the common domain of the φ_i) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$, and such that $\varphi_0(x)$ is minimized.

Problem 6.3. This problem differs from Problem 6.2 only in that \mathfrak{C} and Q_1' are replaced by S_0 and Q_2' , respectively.

It is evident that Problems 6.2 and 6.3 fall under the category of the canonical optimization problem described in [1, §4].

Let z be a solution of Problem 6.2 (or of Problem 6.3), so that (2.3) is satisfied with some $u^* \in \Omega_1$ (or Ω_2). We shall suppose that $u^*(t) \in$ (interior of U) for all $t \in I$ (this requirement can be weakened as indicated in

Note 2.3), and that the vectors $\chi_i', i = 0, 1, \dots, m$, given by (6.1), are linearly independent. It is then easily seen that z is a smoothly regular solution (in the sense of [1, Definition 4.4]) of our canonical optimization problem. Further, it was pointed out in §2 that the sets K_0'' and K_0''' defined by (2.10) and (2.11), respectively, are first-order, convex approximations to $(Q_1' - z) \cap S_0$ and $(Q_2' - z) \cap S_0$, respectively, in the spaces $C \cap S_0$ and S_0 , respectively. If we now appeal to Theorem 4.2 in [1], we conclude that there are numbers $\alpha_0, \dots, \alpha_m$, not all zero, such that $\sum_{i=0}^m \alpha_i \chi_i' \cdot y(t_2) \leq 0$ for all $y \in K_0''$ in the case of Problem 6.2, and all $y \in K_0'''$ in the case of Problem 6.3. Using the definitions of K_0'' and K_0''' given in §2, we can now derive necessary conditions in much the same way that (6.11), (6.6) and (6.7) were obtained. Since the calculations are straightforward and the necessary conditions can be found in [5], we shall omit the remainder of the argument.

The comments in Notes 6.1–6.4 are also pertinent for Problems 6.2 and 6.3.

Note 6.5. In the terminology of control theory, Problem 6.2 deals with sampled-data controls with prescribed “switching times,” and Problem 6.3 deals with controls with free switching times. These problems may be generalized if we reformulate Problem 6.3 by replacing Q_2' by Q_3' (as defined in Note 2.4). Using the first-order, convex approximation indicated in Note 2.4, it is possible to obtain necessary conditions for solutions of this new problem. This new formulation makes it possible to consider sampled-data control systems in which some of the switching times are prescribed and some are free (as in pulse-width modulated systems), and/or in which some of the switching times are not prescribed but are constrained (e.g., the time between two consecutive switchings may be fixed).

Note 6.6. If we take into account the remarks in Note 2.2, and modify the statements of Problems 6.2 and 6.3 accordingly, we can derive necessary conditions for the corresponding optimal control problems. Such problems were first discussed by Chang [6].

The last problem in this section is a so-called discrete optimal control problem.

Let r, k and n be preassigned positive integers with $k > 1$. We shall consider R^{nk} , whose elements we shall represent as k -tuples of vectors in R^n . Let R_0^{nk}, I_0, G and \mathcal{G}_0 be defined as in §4, and let U be a subset of R^r . Let $f(\xi, u, i)$ be a function from $G \times U \times I_0$ into R^n , and let $U_i, i = 1, \dots, k-1$, be subsets of U . We shall suppose that f and the U_i satisfy the hypotheses indicated in Note 4.1. Then let $\tilde{\Omega}$ and $\tilde{\Gamma}$ be defined as in Note 4.1, and let \tilde{Q}' and \tilde{K} be defined in terms of $\tilde{\Gamma}$ as indicated in §4.

Let $\chi_0(\xi), \dots, \chi_m(\xi)$ be real-valued functions, defined for $\xi \in G$, which are of class C^1 , let $W = \{x = (\xi_1, \dots, \xi_k): \xi_i \in R^n \text{ for } i = 1, \dots, k,$

$\xi_1 = 0, \xi_k \in G\}$, so that, since G is open in R^n , W is an open set in R_0^{nk} , and let $\varphi_i, i = 0, \dots, m$, be the continuous functions from W into R^1 defined by the relations

$$(6.12) \quad \varphi_i(\xi_1, \dots, \xi_k) = \chi_i(\xi_k), \quad i = 0, 1, \dots, m.$$

If $z = (\xi_1, \dots, \xi_k) \in W$, then (5.1) is satisfied with $\mathfrak{J} = R_0^{nk}, \varphi = \varphi_i$ and $l = l_i (i = 0, \dots, k)$, where

$$(6.13) \quad l_i(x) = l_i(\xi_1, \dots, \xi_k) = \frac{\partial \chi_i(\xi_k)}{\partial \xi} \cdot \xi_k, \quad i = 0, 1, \dots, m$$

(see the remarks at the beginning of §5).

We can now state our discrete optimal control problem.

Problem 6.4. Let $\mathfrak{J} = R_0^{nk}$ and let $\varphi_0, \varphi_1, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (6.12), where the χ_i have a common open domain $G_0 \subset G$ and are of class C^1 . Then find an element $x \in W \cap \tilde{Q}'$ (where $W \subset R_0^{nk}$ is the common domain of the φ_i) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$ and such that $\varphi_0(x)$ is minimized.

It is evident that Problem 6.4 falls under the category of the canonical optimization problem described in [1, §4] (with $\mu = 0$ and $Q' = \tilde{Q}' \cap R_0^{nk}$).

Let $z = (\xi_1, \dots, \xi_k)$ be a solution of this problem, so that (4.1) holds with $g^* = g^{u^*} \in \tilde{\Gamma}$, i.e.,

$$(6.14) \quad \xi_{i+1} = f(\xi_i, u^*(i), i) \quad \text{for } i = 1, \dots, k-1, \quad \xi_1 = 0,$$

where $u^*(i) \in U_i$ for each $i \in I_0$. Let

$$(6.15) \quad \chi_i' = \frac{\partial \chi_i(\xi_k)}{\partial \xi}, \quad i = 0, 1, \dots, m,$$

and let us suppose that the vectors χ_0', \dots, χ_m' are linearly independent.

It is now easy to verify, on the basis of preceding remarks, that z is a smoothly regular solution of our canonical optimization problem (in the sense of [1, Definition 4.4]). It was pointed out in §4 that the set

$$(6.16) \quad \begin{aligned} \bar{K}_0 &= \{x = (\xi_1, \dots, \xi_k) : \xi_i \in R^n \text{ for } i = 1, \dots, k, \xi_1 = 0, \\ \xi_i &= \sum_{j=1}^{i-1} \Phi_{i,j+1}[f(\xi_j, u(j), j) - f(\xi_j, u^*(j), j)] \end{aligned}$$

$$\text{for } i = 2, \dots, k, \quad u \in \bar{\Omega}\}$$

is a first-order, convex approximation to $(\tilde{Q}' - z) \cap R_0^{nk}$ in R_0^{nk} , where

$$(6.17) \quad \Phi_{i,j} = \Phi_{i,j+1} \frac{\partial f(\xi_j, u^*(j), j)}{\partial \xi} \quad \text{for } 1 \leq j \leq i-1 \leq k-1,$$

$$\Phi_{i,i} = \text{the identity for } i = 1, \dots, k.$$

We can now appeal to Theorem 4.2 in [1] and conclude that there exist real numbers α_i , $i = 0, 1, \dots, m$, not all zero, such that (see (6.13) and (6.15))

$$(6.18) \quad \sum_{i=0}^m \alpha_i \chi_i' \cdot \xi_k \leq 0 \quad \text{whenever} \quad (\xi_1, \dots, \xi_k) \in \tilde{K}_0, \quad \alpha_0 \leq 0.$$

For any $s \in I_0$ and $v \in U_s$, let $u_{v,s} \in \tilde{\Omega}$ be defined as follows: $u_{v,s}(s) = v$, $u_{v,s}(j) = u^*(j)$ for $j \neq s$, $j \in I_0$. Let $\tilde{x} = (\tilde{\xi}_1, \dots, \tilde{\xi}_k)$ be the element in \tilde{K}_0 defined by the recurrence relation in (6.16) with $u = u_{v,s}$. Then it follows from (6.18) with $\xi_k = \tilde{\xi}_k$ that

$$(6.19) \quad \sum_{i=0}^m \alpha_i \chi_i' \cdot \Phi_{k,s+1}[f(\zeta_s, v, s) - f(\zeta_s, u^*(s), s)] \leq 0.$$

Note that (6.19) holds for all $s \in I_0$ and all $v \in U_s$. Let $\psi = (\psi_1, \dots, \psi_k) \in R_0^{nk}$ be defined as follows (considering both the ψ_i and χ_i' to be row-vectors in R^n):

$$(6.20) \quad \psi_j = \sum_{i=0}^m \alpha_i \chi_i' \cdot \Phi_{k,j} \quad \text{for} \quad j = 2, \dots, k.$$

It follows from (6.17), (6.19) and (6.20) that

$$(6.21) \quad \psi_{j+1} f(\zeta_j, u^*(j), j) = \max_{v \in U_j} \psi_{j+1} f(\zeta_j, v, j), \quad j = 1, \dots, k-1,$$

$$(6.22) \quad \psi_j = \psi_{j+1} \frac{\partial f(\zeta_j, u^*(j), j)}{\partial \xi}, \quad j = 2, \dots, k-1,$$

$$(6.23) \quad \psi_k = \sum_{i=0}^m \alpha_i \chi_i', \quad \alpha_0 \leq 0.$$

Since the α_i do not all vanish and the χ_i' are linearly independent, $\psi \neq 0$.

Relations (6.21) and (6.22) may be looked upon as a discrete maximum principle (note the analogy with (6.5) and (6.6)); (6.23) (with (6.15)) is a transversality condition.

Thus we have shown that if $z = (\zeta_1, \dots, \zeta_k) \in R_0^{nk}$ is a solution of Problem 6.4 satisfying (6.14) with $u^* \in \tilde{\Omega}$ such that the vectors χ_i' , $i = 0, 1, \dots, m$, defined by (6.15) are linearly independent, then there exists a nonzero vector $\psi = (\psi_1, \dots, \psi_k) \in R_0^{nk}$ such that (6.21)–(6.23) hold for some constants $\alpha_0, \dots, \alpha_m$. This result was first proved, under slightly stronger hypotheses than those made here, by Halkin [20].

Note 6.7. Problem 6.4 is a discrete optimal control problem with fixed left-hand endpoint. Problems with a variable left-hand endpoint (which were considered in [20]) can also be considered by slightly changing the problem statement and reasoning almost exactly as above (there are a

few minor complications). The necessary conditions satisfied by solutions of this problem differ from those for Problem 6.4 only in that there is an additional transversality condition. (See Note 6.1.) We can also make remarks for Problem 6.4 which are analogous to Notes 6.2–6.4.

7. Applications to optimization problems which arise from quasiconvex families. In this section we shall consider a general variational problem which includes, as special cases, Problem 6.1 together with the generalizations indicated in Notes 6.1 and 6.2, as well as problems with “free” initial and terminal times.

Let G again be a nonempty open set in R^n , and let $I = [t_1, t_2]$ be a fixed compact interval. Let \mathcal{J} be defined as in §3 and \mathcal{C} as in §2, let Γ be a given quasiconvex family of functions in \mathcal{J} , and let Q^{**} be correspondingly defined as indicated in §3. Then consider the following problem.

Problem 7.1. Let $\mathfrak{J} = \mathcal{C} \times R^k$, let D be a convex subset of $I^k \subset R^k$, and let $\varphi_{-\mu}, \dots, \varphi_{-1}, \varphi_0, \varphi_1, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (5.6), where the corresponding real-valued functions θ_i are of class C^1 and have a common domain \tilde{G} . We suppose that \tilde{G} is an open set in $R^{nk} \times I^k$ whose intersection with $G^k \times D$ is not empty. Then find an element $\mathbf{x} = (x, \tau_1, \dots, \tau_k) \in W_1 \cap (Q^{**} \times D)$ (where $W_1 \subset \mathfrak{J}$ is defined by (5.4) and (5.5)) such that $\varphi_i(\mathbf{x}) = 0$ for $i = 1, \dots, m$, $\varphi_{-i}(\mathbf{x}) \leq 0$ for $i = 1, \dots, \mu$, and such that $\varphi_0(\mathbf{x})$ is minimized.

Note 7.1. A more natural problem than Problem 7.1 is the following. Instead of requiring that $\mathbf{x} = (x, \tau_1, \dots, \tau_k) \in W_1 \cap (Q^{**} \times D)$ —so that $x(t)$ must be defined on all of I and must satisfy the equation $\dot{x}(t) = g(x(t), t)$ for almost all $t \in I$ (and some $g \in \Gamma$)—require only that $x(t)$ be defined, be absolutely continuous, and take on its values in G for $t \in [\tau', \tau'']$, where $\tau' = \min_{1 \leq i \leq k} \tau_i$, $\tau'' = \max_{1 \leq i \leq k} \tau_i$, that $\dot{x}(t) = g(x(t), t)$ for almost all $t \in [\tau', \tau'']$ (and some $g \in \Gamma$), that $(\tau_1, \dots, \tau_k) \in D$ and that $(x(\tau_1), \dots, x(\tau_k), \tau_1, \dots, \tau_k) \in \tilde{G}$. Let $\mathbf{z} = (z, \tau_1^*, \dots, \tau_k^*)$ be a solution of the modified problem, so that

$$(7.1) \quad \dot{z}(t) = g^*(z(t), t)$$

for almost all $t \in [\tau^{**}, \tau^{***}]$ and some $g^* \in \Gamma$ (where $\tau^{**} = \min \tau_i^*$, $\tau^{***} = \max \tau_i^*$). Since G is open, the solution of (7.1) can be defined, and will take on its values in G , for $\tau^{**} - \epsilon_1 \leq t \leq \tau^{***} + \epsilon_2$ where $\epsilon_1 > 0$ if $\tau^{**} > t_1$ and $\epsilon_1 = 0$ if $\tau^{**} = t_1$, and similarly for ϵ_2 . Further, \mathbf{z} is now a solution of a problem which differs from Problem 7.1 only in that I is replaced by $I' = [\tau^{**} - \epsilon_1, \tau^{***} + \epsilon_2]$, and D by $(D \cap I'^k)$. But the necessary conditions which will be derived for solutions of Problem 7.1 are unchanged if, in the problem statement, I and D are replaced by I' and $D \cap I'^k$, respectively. Consequently solutions of the “more natural problem” satisfy the same necessary conditions derived below as the solutions of Problem 7.1.

Note 7.2. It is easily seen that Problem 6.1, as well as the generalizations indicated in Notes 6.1 and 6.2, are special cases of Problem 7.1 (see Note 3.1). Further, the necessary conditions for solutions of Problem 7.1 which we shall derive will imply the conditions derived in §6 for Problem 6.1. This brings up the natural question: Why was Problem 6.1 considered separately? There are two reasons. The first is historical: the derivation presented in §6, together with the associated arguments in §2, are in essence the same as those originally presented by Pontryagin et al. [2], [7]. The second reason is that some of the “variations” used in §§2 and 6 for Problem 6.1 are required in the treatment of Problem 6.3 (as well as the extensions described in Notes 6.5 and 6.6), which is *not* a particular case of Problem 7.1.

For the sake of definiteness, and for ease of notation, we shall confine ourselves to the case of Problem 7.1 wherein $k = 3$ and $D = \{(\tau_1, \tau_2, \tau_3): t_1 \leq \tau_1 < \tau_3 < \tau_2 \leq t_2\}$. This case includes the most important features of the general problem, and the arguments used for this special case can obviously be extended to the general case.

It is clear that Problem 7.1 falls under the category of the canonical optimization Problem of [1, §4] (with $Q' = Q^* \times D$).

Let $\mathbf{z} = (z, \tau_1^*, \tau_2^*, \tau_3^*)$ be a solution of Problem 7.1, so that $t_1 \leq \tau_1^* < \tau_3^* < \tau_2^* \leq t_2$, and (3.1) holds. Let us suppose that, for each $i = 1, 2, 3$, $z(t)$ is differentiable at $t = \tau_i^*$ and $\dot{z}(\tau_i^*) = g^*(z(\tau_i^*), \tau_i^*)$. Also suppose that $t_1 < \tau_1^* < \tau_2^* < t_2$. We shall use the notations \mathcal{G}_z and \mathcal{J}_z with the same meaning as in [1, §4]. Let

$$(7.2) \quad \begin{aligned} \theta_i^{(j)} &= \frac{\partial \theta_i(z(\tau_1^*), z(\tau_2^*), z(\tau_3^*), \tau_1^*, \tau_2^*, \tau_3^*)}{\partial \xi_j}, \\ \theta_i^{(3+j)} &= \frac{\partial \theta_i(z(\tau_1^*), z(\tau_2^*), z(\tau_3^*), \tau_1^*, \tau_2^*, \tau_3^*)}{\partial \sigma_j}, \\ i &= -\mu, \dots, m, \quad j = 1, 2, 3. \end{aligned}$$

Our final assumption is that the relations

$$(7.3) \quad \sum_{i=-\mu}^m \alpha_i(\theta_i', \theta_i'', \theta_i''', \theta_i^{(iv)}, \theta_i^{(v)}, \theta_i^{(vi)}) = 0, \\ \alpha_{-i} \leq 0 \quad \text{for } i \in \mathcal{G}_z, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z,$$

imply that $\alpha_i = 0$ for all $i = -\mu, \dots, m$. It now follows from (5.7) and (5.1) that \mathbf{z} is a smoothly regular solution of our canonical optimization problem (in the sense of [1, Definition 4.4]).

In §3 it was shown that K^* defined by (3.2)–(3.4) is a first-order, convex approximation to $(Q^{*'} - z)$ in \mathcal{C} (see Theorem 3.1). Consequently, if

$$(7.4) \quad K = K^* \times (D - (\tau_1^*, \tau_2^*, \tau_3^*)),$$

then K is a first-order, convex approximation to $[(Q^{*'} \times D) - \mathbf{z}]$ (see [1, Notes 2.2 and 2.3]). It now follows from Theorem 4.2 in [1] that there exist real numbers α_i , $i = -\mu, \dots, m$, such that

$$(7.5) \quad \sum_{i=-\mu}^m \alpha_i l_i(x, \tau_1, \tau_2, \tau_3) \leq 0 \quad \text{for all } (x, \tau_1, \tau_2, \tau_3) \in K,$$

$$\sum_{i=-\mu}^m |\alpha_i| > 0, \quad \alpha_{-i} \leq 0 \quad \text{for } i \in \mathcal{J}_z, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z,$$

where (see (5.7) and (7.2))

$$(7.6) \quad l_i(x, \tau_1, \tau_2, \tau_3) = \sum_{j=1}^3 \{\theta_i^{(j)} \cdot x(\tau_j^*) + \tau_j [\theta_i^{(j)} \cdot \dot{z}(\tau_j^*) + \theta_i^{(3+j)}]\},$$

$$i = -\mu, \dots, m.$$

Since $(\delta x_{\xi, g^*}, 0, 0, 0) \in K$ for all $\xi \in R^n$, it follows from (7.5) and (7.6), by virtue of (3.2), that

$$(7.7) \quad \sum_{i=-\mu}^m \sum_{j=1}^3 \alpha_i \theta_i^{(j)} \cdot \Phi(\tau_j^*) \xi \leq 0 \quad \text{for all } \xi \in R^n,$$

which is possible only if (considering the $\theta_i^{(j)}$ to be row vectors)

$$(7.8) \quad \sum_{i=-\mu}^m \sum_{j=1}^3 \alpha_i \theta_i^{(j)} \Phi(\tau_j^*) = 0.$$

Also, $(\delta x_{0, g}, 0, 0, 0) \in K$ for all $g \in \Gamma$, so that, by the same token,

$$(7.9) \quad \sum_{i=-\mu}^m \sum_{j=1}^3 \alpha_i \theta_i^{(j)} \Phi(\tau_j^*) \int_{t_1}^{\tau_j^*} \Phi^{-1}(\tau) [g(z(\tau), \tau) - g^*(z(\tau), \tau)] d\tau \leq 0$$

for all $g \in \Gamma$.

In (7.7)–(7.9), Φ denotes the (absolutely continuous) solution of (3.3). Let $\psi_j(t) = \sum_{i=-\mu}^m \alpha_i \theta_i^{(j)} \Phi(\tau_j^*) \Phi^{-1}(t)$, $j = 1, 2, 3$, so that the $\psi_j(t)$ are absolutely continuous, row-vector functions that satisfy the relations

$$(7.10) \quad \dot{\psi}_j(t) = -\psi_j(t) g_x^*(z(t), t) \quad \text{for almost all } t \in I,$$

$$(7.11) \quad \psi_j(\tau_j^*) = \sum_{i=-\mu}^m \alpha_i \theta_i^{(j)},$$

$$(7.12) \quad \sum_{j=1}^3 \psi_j(t) = 0 \quad \text{for all } t \in I.$$

(Relation (7.12) is a consequence of (7.8).) Hence, (7.9) can be rewritten in the form

$\psi(t)$ defined on I such that (7.14) and (7.16)–(7.22) hold for some real numbers $\alpha_{-\mu}, \dots, \alpha_m$. Relation (7.14) is a maximum condition, (7.18)–(7.22) with (7.2) represent transversality conditions. These results were first obtained by Gamkrelidze [3].

Note 7.3. It is possible to remove the hypotheses that $\tau_1^* > t_1$ and $\tau_2^* < t_2$ in the analysis of Problem 7.1. However, one must then employ a theorem which is slightly sharper than Theorem 4.2 in [1]. The necessary conditions obtained under the weaker hypotheses coincide with those obtained for Problem 7.1 except that the equality in (7.21) ((7.22)) must be replaced by \geq if $\tau_1^* = t_1$ ($\tau_2^* = t_2$). In this case, however, it may turn out that $\psi(t) \equiv 0$ on I , so that the principal necessary condition (7.14) is trivially satisfied. If one makes an assumption related to, but stronger than, the one made with relation to (7.3) in the case of Problem 7.1, one can also conclude that $\psi(t) \not\equiv 0$ on I in this case.

Note 7.4. A special case of Problem 7.1 is one in which the “times” τ_i , $i = 1, 2, 3$, are all fixed, i.e., among the functionals φ_i with $i > 0$, there are functionals of the form $\varphi_i(x, \tau_1, \dots, \tau_k) = \tau_j - \tau_j^*$ (where $\tau_j^* \in I$ is fixed) for $j = 1, 2$, and 3. It is then more convenient to reformulate Problem 7.1 by considering that the underlying space \mathfrak{J} is \mathfrak{C} (rather than $\mathfrak{C} \times R^k$), and that the functionals φ_i are of the form of (5.2), with $\tau_i = \tau_i^*$, so that the differentials l_i are given by (5.3). It is then possible to derive necessary conditions *without having to assume that $z(t)$ is differentiable at $t = \tau_i^*$* for $i = 1, 2$, or 3. These conditions, which can be obtained by arguments virtually unchanged from those presented above, are essentially identical with those derived for solutions of the general Problem 7.1, except that (7.20)–(7.22) must be omitted. Also, every solution \mathbf{z} of the general Problem 7.1 is a fortiori a solution of a “fixed-time” problem, so that, even if z is not differentiable at τ_1^* , τ_2^* , and τ_3^* (but the other hypotheses are satisfied), the above necessary conditions—with the exception of (7.20)–(7.22)—hold. If $z(t)$ is differentiable at one or two of the points τ_i^* , then the corresponding relation of (7.20)–(7.22) is satisfied.

Note 7.5. The form of the necessary conditions (7.14) and (7.16)–(7.22) will basically be unchanged if the roles of some of the φ_i in the problem statement are interchanged (see [1, Note 4.4]).

Note 7.6. Without going into detail, we remark that comments similar to those made in Note 6.4 can also be made here.

Let us now suppose that the quasiconvex class Γ in Problem 7.1 is the class Γ^* described in Note 3.1. Let us suppose in addition that the corresponding function $f(x, u, t)$ is continuous in all of its arguments, and that $U_t = U$ for every t . Let $\mathbf{z}^* = (z, \tau_1^*, \tau_2^*, \tau_3^*)$ be a solution of this problem satisfying the hypotheses indicated previously. Thus z satisfies (3.1), where

$$(7.23) \quad g^*(x, t) = f(x, u^*(t), t), \quad u^* \in \Omega^*$$

(Ω^* is defined as in Note 3.1). Let I_1 denote the set of $t \in (t_1, t_2)$ for which u^* is regular (see (2.4)) and $u^*(t) \in U$. Then $\text{meas } I_1 = \text{meas } I$. As before, we can show that there exist real numbers $\alpha_i, i = -\mu, \dots, m$, such that (7.4)–(7.6) hold. Let us now imbed $\mathfrak{J} = \mathbb{C} \times R^k$ in $\mathfrak{S} \times R^k$, where \mathfrak{S} is defined as in §2. It is easily verified that functionals l of the form (5.7) are defined, linear, and continuous in $\mathfrak{S} \times R^k$. It then follows from (7.5) that

$$(7.24) \quad \sum_{i=-\mu}^m \alpha_i l_i(x, \tau_1, \tau_2, \tau_3) \leq 0 \quad \text{for all } (x, \tau_1, \tau_2, \tau_3) \in \overline{\text{cone } K},$$

where $\overline{\text{cone } K}$ denotes the closure in $\mathfrak{S} \times R^k$ of $(\text{cone } K) = \{\eta \mathbf{x} : \mathbf{x} \in K, \eta \geq 0\}$.

For every $s \in I_1$ and $v \in U$, define the function $x_{s,v}(t) \in \mathfrak{S}$ as follows:

$$(7.25) \quad x_{s,v}(t) = \begin{cases} 0 & \text{for } t_1 \leq t < s, \\ \Phi(t)\Phi^{-1}(s)[f(z(s), v, s) - f(z(s), u^*(s), s)] & \text{for } s \leq t \leq t_2, \end{cases}$$

where $\Phi(t)$ is the absolutely continuous $n \times n$ matrix function that satisfies (3.3) (with g^* given by (7.23)). Obviously, $\dot{x}_{s,v}(t) = f_x(z(t), u^*(t), t) x_{s,v}(t)$ for almost all $t \in I$. Let us show that $(x_{s,v}, 0, 0, 0) \in \overline{\text{cone } K}$ whenever $(s, v) \in I_1 \times U$. Indeed, it is evident that if $s \in I_1, v \in U, i > 0$ and

$$(7.26) \quad g_i(x, t) = \begin{cases} f(x, v, t) & \text{for } s - \frac{1}{i} \leq t < s, \\ f(x, u^*(t), t) & \text{for } t_1 \leq t < s - \frac{1}{i} \text{ and } s \leq t \leq t_2, \end{cases}$$

then $g_i(x, t) \in \Gamma^*$ whenever i is sufficiently large. But if $g_i \in \Gamma^*$, then $(i\delta x_{0,g_i}, 0, 0, 0) \in \text{cone } K$ (see (3.4) and (7.4)). Since $s \in I_1, i\delta x_{0,g_i} \rightarrow x_{s,v}$ in \mathfrak{S} as $i \rightarrow \infty$. Consequently, $(x_{s,v}, 0, 0, 0) \in \overline{\text{cone } K}$ whenever $s \in I_1$ and $v \in U$. If $\tau_1^* < s < \tau_3^*$, it follows from (7.25) that

$$(7.27) \quad \begin{aligned} x_{s,v}(\tau_1^*) &= 0, \\ x_{s,v}(\tau_j^*) &= \Phi(\tau_j^*)\Phi^{-1}(s)[f(z(s), v, s) - f(z(s), u^*(s), s)] \end{aligned} \quad \text{for } j = 2 \text{ or } 3;$$

if $\tau_3^* < s < \tau_2^*$, then

$$(7.28) \quad \begin{aligned} x_{s,v}(\tau_j^*) &= 0 \quad \text{for } j = 1 \text{ and } 3, \\ x_{s,v}(\tau_2^*) &= \Phi(\tau_2^*)\Phi^{-1}(s)[f(z(s), v, s) - f(z(s), u^*(s), s)]. \end{aligned}$$

As before, let $\psi_j(t) = \sum_{i=-\mu}^m \alpha_i \theta_i^{(j)} \Phi(\tau_j^*)\Phi^{-1}(t)$ for $j = 1, 2, 3$, and let $\psi(t)$ be given by (7.15). It then follows from (7.24), (7.6), (7.12) (which

follow as before), (7.27) and (7.28) that

$$(7.29) \quad \psi(t)f(z(t), u^*(t), t) = \max_{v \in U} \psi(t)f(z(t), v, t) \\ \text{for almost all } t \in [\tau_1^*, \tau_2^*],$$

(recall that $\text{meas } I_1 = \text{meas } I$).

It is easily seen that (7.16)–(7.22) follow as before (where g^* is given by (7.23)).

In summary, if $\mathbf{z} = (z, \tau_1^*, \tau_2^*, \tau_3^*)$ is a solution of Problem 7.1 where $\Gamma = \Gamma^*$ (as defined in Note 3.1), and all the hypotheses indicated above are satisfied, and if z satisfies (3.1) and (7.23), then there exists a non-trivial (row) n -vector valued function $\psi(t)$ defined on I such that (7.16)–(7.23) and (7.29) hold for some real numbers $\alpha_{-\mu}, \dots, \alpha_m$. This result was first proved in [3], and was also obtained independently by Pshenichniy [21] for the special case where $f(x, u, t)$ has the form $A(t)x + G(t, u)$. The proofs in [21] are very much in the spirit of those used in the present work. Note that (7.29) is the conventional maximum principle.

Note 7.7. In the preceding special case of Problem 7.1, the same results can be derived even if the assumption that $U_t = U$ for all t is considerably weakened. Weaker hypotheses of this kind were discussed by Guinn [8], who was the first to obtain a “maximum principle” under such mild assumptions.

Note 7.8. The remarks of Notes 7.3–7.6 are here applicable also.

Note 7.9. It is possible to define a problem which generalizes Problem 6.4 in the same manner that Problem 7.1 generalizes Problem 6.1. Using the first-order, convex approximation described in §4, it is an easy matter to obtain necessary conditions for solutions of this generalized “discrete” optimization problem.

8. Applications to optimization problems with restricted phase coordinates and to minimax problems. In this section we shall discuss two problems whose solutions are generally totally regular, but not smoothly regular (in the sense of [4, Definitions 4.3 and 4.4]).

The first of these problems which we shall consider is commonly referred to as an optimization problem with restricted phase coordinates.

Let G, I, \mathcal{g} and \mathcal{C} be defined as before, let Γ be a given quasiconvex family of functions in \mathcal{g} , and let Q^{**} be correspondingly defined as in §3.

Problem 8.1. Let $\mathfrak{J} = \mathcal{C}$, and let $\varphi_{-\mu}, \dots, \varphi_0, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (5.2), where the corresponding real-valued functions χ_i are of class C^1 and have a common open domain $G_0 \subset G \times G, k = 2, \tau_1 = t_1$, and $\tau_2 = t_2$. Let $\varphi_{-\mu-1}$ be a functional on \mathfrak{J} of the form of (5.9), where \tilde{g} is a continuous function from $G \times I$ into \mathbb{R}^1 such that \tilde{g}_x is defined and con-

tinuous on $G \times I$, and \tilde{I} is a fixed subset of I . Then find an element $x \in W \cap Q^{*'} (where $W \subset \mathfrak{J}$ is the common domain of $\varphi_{-\mu-1}, \dots, \varphi_m$) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$, and $\varphi_{-i}(x) \leq 0$ for $i = 1, \dots, \mu + 1$, and such that $\varphi_0(x)$ is minimized.$

If we set $Q' = Q^{*'}$, Problem 8.1 is a particular case of the canonical optimization problem described in [1, §4].

Let z be a solution thereof, so that (3.1) is satisfied.

If $\varphi_{-\mu-1}(z) < 0$, then z is a local solution of the canonical optimization obtained from Problem 8.1 by deleting $\varphi_{-\mu-1}$, and this latter problem is a special case of Problem 7.1 (see Note 7.4). Consequently, we shall suppose that $\varphi_{-\mu-1}(z) = 0$. Let I_e be defined by (5.11) and (5.9). We shall also suppose that $\tilde{g}_x(z(t), t) \neq 0$ for all $t \in I_e$.

Let

$$(8.1) \quad \begin{aligned} \chi_i' &= \frac{\partial \chi_i(z(t_1), z(t_2))}{\partial \xi_1}, & \chi_i'' &= \frac{\partial \chi_i(z(t_1), z(t_2))}{\partial \xi_2}, \\ i &= -\mu, \dots, 0, \dots, m, \\ \tilde{g}' &= \tilde{g}_x(z(t_1), t_1), & \tilde{g}'' &= \tilde{g}_x(z(t_2), t_2). \end{aligned}$$

If $t_i \in I_e$ for $i = 1$ and 2 , we shall assume that the relations

$$(8.2) \quad \alpha_{-\mu-1}\tilde{g}' + \sum_{i=-\mu}^m \alpha_i \chi_i' = \alpha_{-\mu-2}\tilde{g}'' + \sum_{i=-\mu}^m \alpha_i \chi_i'' = 0,$$

$$\alpha_{-i} \leq 0 \quad \text{for } i \geq 0, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z$$

(where \mathcal{J}_z is defined as in [1, §4]) imply that $\alpha_i = 0$ for all $i = -\mu - 2, \dots, m$. If $t_1 \in I_e$, $t_2 \notin I_e$, we shall suppose that (8.2) together with the relation $\alpha_{-\mu-2} = 0$ imply that $\alpha_i = 0$ for all i , and the obvious analog of this assumption is made if $t_2 \in I_e$, $t_1 \notin I_e$. Finally, if $t_1 \notin I_e$ and $t_2 \notin I_e$, we shall suppose that (8.2), together with $\alpha_{-\mu-1} = \alpha_{-\mu-2} = 0$, imply that $\alpha_i = 0$ for all i .

It is now straightforward to verify, on the basis of the results of §5, that z is a totally regular local solution of our canonical optimization problem in the sense of [1, Definition 4.3]. Since K^* (see (3.2)–(3.4)) is a first-order, convex approximation to $Q^{*'} - z$ in \mathfrak{C} (see Theorem 3.1), we can conclude, on the basis of Theorem 4.2 in [1], that there exist real numbers $\alpha_{-\mu-1}, \dots, \alpha_0, \dots, \alpha_m$, not all zero, such that

$$(8.3) \quad \sum_{i=-\mu}^m \alpha_i l_i(x) + \alpha_{-\mu-1} c(x) \leq 0 \quad \text{for all } x \in K^*,$$

$$\alpha_{-i} \leq 0 \quad \text{for } i \geq 0, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z,$$

where $c(x)$ is given by (5.10) and (see (5.3) and (8.1))

$$(8.4) \quad l_i(x) = \chi_i' \cdot x(t_1) + \chi_i'' \cdot x(t_2) \quad \text{for } i = -\mu, \dots, m.$$

Since $\delta x_{\xi, g}^* \in K^*$ for all $\xi \in R^n$, it follows from (8.3), (8.4) and (5.10), by virtue of (3.2) and (3.3), that

$$(8.5) \quad \sum_{i=-\mu}^m \alpha_i [\chi_i' + \chi_i'' \Phi(t_2)] \xi \leq \sup_{t \in I_e} [-\alpha_{-\mu-1} \tilde{g}_x(z(t), t) \Phi(t) \xi] \quad \text{for all } \xi \in R^n$$

(in (8.5), and henceforth, χ_i' , χ_i'' , and \tilde{g}_x are to be considered row-vectors). It follows from (8.5) and the compactness of the set $\{\tilde{g}_x(z(t), t) \Phi(t) : t \in I_e\}$ that

$$(8.6) \quad \sum_{i=-\mu}^m \alpha_i [\chi_i' + \chi_i'' \Phi(t_2)] \in \text{co} \{ -\alpha_{-\mu-1} \tilde{g}_x(z(t), t) \Phi(t) : t \in I_e \},$$

where $\text{co } E$ denotes the convex hull of E .

Further, $\delta x_{0, g} \in K^*$ whenever $g \in \Gamma$, so that, if we set

$$(8.7) \quad \psi(\tau) = \sum_{i=-\mu}^m \alpha_i \chi_i'' \Phi(t_2) \Phi^{-1}(\tau), \quad \tau \in I,$$

$$(8.8) \quad \psi_1(\tau, t) = \tilde{g}_x(z(t), t) \Phi(t) \Phi^{-1}(\tau), \quad t \in I, \quad \tau \in I,$$

we obtain (by virtue of (8.3), (8.4), (5.10), (3.2) and (3.3))

$$(8.9) \quad \int_{t_1}^{t_2} \psi(\tau) [g(z(\tau), \tau) - g^*(z(\tau), \tau)] d\tau + \alpha_{-\mu-1} \sup_{t \in I_e} \int_{t_1}^t \psi_1(\tau, t) [g(z(\tau), \tau) - g^*(z(\tau), \tau)] d\tau \leq 0$$

for all $g \in \Gamma$,

$$(8.10) \quad \frac{d\psi(\tau)}{d\tau} = -\psi(\tau) g_x^*(z(\tau), \tau) \quad \text{for almost all } \tau \in I,$$

$$(8.11) \quad \frac{\partial \psi_1(\tau, t)}{\partial \tau} = -\psi_1(\tau, t) g_x^*(z(\tau), \tau)$$

for almost all $\tau \in I$ and every $t \in I$,

$$(8.12) \quad \psi(t_2) = \sum_{i=-\mu}^m \alpha_i \chi_i'',$$

$\alpha_{-i} \leq 0$ for $i \geq 0$, $\alpha_{-i} = 0$ for $i \in \mathcal{J}_z$,

$$(8.13) \quad \psi_1(t, t) = \tilde{g}_x(z(t), t) \quad \text{for all } t \in I,$$

and, by virtue of (8.6) and (3.3),

$$(8.14) \quad \psi(t_1) = -\sum_{i=-\mu}^m \alpha_i \chi_i' + \tilde{\chi},$$

where

$$(8.15) \quad \tilde{\chi} \in \text{co} \{ -\alpha_{-\mu-1} \tilde{g}_x(z(t), t) \Phi(t) : t \in I_e \},$$

$$\alpha_{-\mu-1} \leq 0, \quad \sum_{i=-\mu-1}^m |\alpha_i| > 0.$$

Thus, we have shown that if z is a solution of Problem 8.1 satisfying the hypotheses indicated above, and if (3.1) holds, then there exist absolutely continuous, (row) n -vector valued functions (of $\tau \in I$) $\psi(\tau)$ and $\psi_1(\tau, t)$, the latter being defined for each $t \in I$, such that (8.9)–(8.15) are satisfied for some real constants $\alpha_i, i = -\mu - 1, \dots, m$, where I_e is given by (5.11) and (5.9). Relation (8.9) is the maximum condition; (8.12)–(8.15), with (8.1), play the role of transversality conditions. These necessary conditions were first obtained by Gamkrelidze [9].

If the quasiconvex class Γ in Problem 8.1 is the class Γ^* described in Note 3.1, where the corresponding function $f(x, u, t)$ is continuous in all of its arguments, and $U_t = U$ for each $t \in I$, then (8.9) may be rewritten in a different form. Namely, suppose that (7.23) holds; and denote by I_1 the set of points $t \in (t_1, t_2)$ for which u^* is regular (see (2.4)) and $u^*(t) \in U$. Let $s \in I_1$ and $v \in U$ be arbitrary, and let functions $g_i(x, t)$ be defined by (7.26), so that $g_i \in \Gamma^*$ whenever i is sufficiently large. Let

$$(8.16) \quad h(\tau) = f(z(\tau), v, \tau) - f(z(\tau), u^*(\tau), \tau).$$

Also, let $t_1^* = \min_{t \in I_e} t$ and $t_2^* = \max_{t \in I_e} t$. Setting $g = g_i$ in (8.9), we obtain (for i sufficiently large)

$$(8.17) \quad \int_{s-(1/i)}^s \psi(\tau) h(\tau) d\tau \leq 0 \quad \text{if } t_2^* < s < t_2,$$

$$(8.18) \quad \int_{s-(1/i)}^s \psi(\tau) h(\tau) d\tau + \alpha_{-\mu-1} \sup_{t \in I_e} \int_{s-(1/i)}^s \psi_1(\tau, t) h(\tau) d\tau \leq 0$$

if $t_1 < s \leq t_1^*$,

$$(8.19) \quad \int_{s-(1/i)}^s \psi(\tau) h(\tau) d\tau + \alpha_{-\mu-1} \sup_{\substack{t \in I_e \\ t > s-(1/i)}}^+ \int_{s-(1/i)}^{\min\{s, t\}} \psi_1(\tau, t) h(\tau) d\tau \leq 0$$

if $t_1^* < s \leq t_2^*$,

where $\sup^+ = \sup$ if $\sup \geq 0$ and $\sup^+ = 0$ if $\sup \leq 0$. Since s is a regular point for u^* , and hence also for ψh ,

$$(8.20) \quad i \int_{s-(1/i)}^s \psi(\tau) h(\tau) d\tau \xrightarrow{i \rightarrow \infty} \psi(s) h(s).$$

Further, it is not difficult to verify that

$$(8.21) \quad i \sup_{t \in I_e} \int_{s-(1/i)}^s \psi_1(\tau, t) h(\tau) d\tau \xrightarrow{i \rightarrow \infty} \sup_{t \in I_e} \psi_1(s, t) h(s) \quad \text{if } t_1 < s \leq t_1^*,$$

$$(8.22) \quad i \sup_{\substack{t \in I_e \\ t > s-(1/i)}}^+ \int_{s-(1/i)}^{\min\{s, t\}} \psi_1(\tau, t) h(\tau) d\tau \xrightarrow{i \rightarrow \infty} \sup_{\substack{t \in I_e \\ t \geq s}}^+ \psi_1(s, t) h(s) \\ \text{if } t_1^* < s \leq t_2^*.$$

It now follows from (8.17)–(8.22) that

$$(8.23) \quad \begin{aligned} \psi(s)h(s) + \alpha_{-\mu-1} \sup_{t \in I_e} \psi_1(s, t)h(s) &\leq 0 \quad \text{if } t_1 < s \leq t_1^*, \\ \psi(s)h(s) + \alpha_{-\mu-1} \sup_{\substack{t \in I_e \\ t \geq s}}^+ \psi_1(s, t)h(s) &\leq 0 \quad \text{if } t_1^* < s \leq t_2^*, \\ \psi(s)h(s) &\leq 0 \quad \text{if } t_2^* < s < t_2. \end{aligned}$$

Now the appropriate one of (8.23) must hold whenever $s \in I_1$ (i.e., for almost all $s \in I$), whatever $v \in U$. Thus (see (8.16)), since $\text{meas } I_1 = \text{meas } I$,

$$(8.24) \quad \begin{aligned} &\max_{v \in U} \{\psi(s)f(z(s), v, s) \\ &\quad + \alpha_{-\mu-1} \sup_{t \in I_e} \psi_1(s, t) [f(z(s), v, s) - f(z(s), u^*(s), s)]\} \\ &= \psi(s)f(z(s), u^*(s), s) \quad \text{for almost all } s \in [t_1, t_1^*], \\ &\max_{v \in U} \{\psi(s)f(z(s), v, s) \end{aligned}$$

$$(8.25) \quad \begin{aligned} &\quad + \alpha_{-\mu-1} \sup_{\substack{t \in I_e \\ t \geq s}}^+ \psi_1(s, t) [f(z(s), v, s) - f(z(s), u^*(s), s)]\} \\ &= \psi(s)f(z(s), u^*(s), s) \quad \text{for almost all } s \in [t_1^*, t_2^*], \end{aligned}$$

$$(8.26) \quad \begin{aligned} &\max_{v \in U} \psi(s)f(z(s), v, s) \\ &= \psi(s)f(z(s), u^*(s), s) \quad \text{for almost all } s \in [t_2^*, t_2], \end{aligned}$$

$$(8.27) \quad \alpha_{-\mu-1} \leq 0, \quad t_1^* = \min_{t \in I_e} t, \quad t_2^* = \max_{t \in I_e} t.$$

Relations (8.24)–(8.27) are the maximum condition for the special case of Problem 8.1 under consideration. Relation (8.26) is evidently of the same form as (7.29), but for $s < t_2^*$, the quantity being maximized contains an extra term, which, it is important to note, generally depends (through $\psi_1(s, t)$ —see (8.11) and (8.13)) on values of $z(t)$ for $t > s$. Conditions (8.24)–(8.27) were also first pointed out by Gamkrelidze [9].

Let us now return to the more general form of Problem 8.1 and derive an alternate set of necessary conditions satisfied by solutions thereof. Thus, let z be a solution of Problem 8.1 satisfying the conditions indicated earlier, so that z is a totally regular local solution of the corresponding canonical optimization problem. Let us appeal to Theorem 4.1 in [1], taking into account [1, Lemma 4.4] and the fact that K^* defined by (3.1)–(3.4) is a first-order, convex approximation to $Q^{*'} - z$ in \mathcal{C} . Consequently, there exist numbers $\alpha_{-\mu}, \dots, \alpha_0, \dots, \alpha_m$, not all zero, and a functional $l_{-\mu-1} \in \mathcal{C}^*$, such that

$$(8.28) \quad \sum_{i=-\mu}^m \alpha_i l_i(x) + l_{-\mu-1}(x) \leq 0 \quad \text{for all } x \in K^*,$$

$$\alpha_{-i} \leq 0 \quad \text{for } i \geq 0, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z,$$

where the l_i , for $i = -\mu, \dots, m$, are given by (8.4) and (8.1), and (see (5.10))

$$(8.29) \quad l_{-\mu-1}(y) \geq 0 \quad \text{whenever} \quad \sup_{t \in I_e} [\tilde{g}_x(z(t), t) \cdot y(t)] < 0 \quad \text{and} \quad y \in \mathcal{C}.$$

Also, it is easy to show, by virtue of our assumptions regarding (8.2), that $Z' \neq \{0\}$ (Z' is defined in [1, Theorem 4.1]); consequently,

$$(8.30) \quad \sum_{i=-\mu}^m \alpha_i l_i + l_{-\mu-1} \neq 0.$$

We shall prove that (8.29) implies that there exists a scalar-valued, non-increasing function $\lambda(t)$, defined on I and continuous from the right in (t_1, t_2) , such that

$$(8.31) \quad \lambda(t_2) = 0, \quad \lambda \text{ is constant on every subinterval of } I$$

which does not meet I_e ,

$$(8.32) \quad l_{-\mu-1}(y) = \int_{t_1}^{t_2} [\tilde{g}_x(z(t), t) \cdot y(t)] d\lambda(t) \quad \text{for all } y \in \mathcal{C}.$$

To show the existence of such a function λ , we first prove the following lemma, which is a generalization of the well-known Farkas-Minkowski lemma.

LEMMA 8.1. *Let L be a continuous linear mapping from a real Banach space \mathcal{B}_1 onto a real Banach space \mathcal{B}_2 , let Z_2 be a convex cone in \mathcal{B}_2 with vertex at 0 such that Z_2 has a nonempty interior and $Z_2 \neq \mathcal{B}_2$, and let $Z_1 = L^{-1}(Z_2)$. Then Z_1 is a convex cone in \mathcal{B}_1 with vertex at 0 and a nonempty interior, and $Z_1 \neq \mathcal{B}_1$. Further if $l' \in \mathcal{B}_1^*$ and $l'(y) \leq 0$ for all $y \in Z_1$, then there exists a functional $l'' \in \mathcal{B}_2^*$ such that $l''(w) \leq 0$ for all $w \in Z_2$ and $l'(y) = l''(Ly)$ for all $y \in \mathcal{B}_1$ (\mathcal{B}_i^* denotes the conjugate space of \mathcal{B}_i).*

Proof. Since L is linear, continuous, and onto \mathfrak{B}_2 , the hypotheses on Z_2 immediately imply that Z_1 is a convex cone in \mathfrak{B}_1 with vertex at 0 and a non-empty interior, and that $Z_1 \neq \mathfrak{B}_1$. Let $l' \in \mathfrak{B}_1^*$ be such that $l'(y) \leq 0$ for all $y \in Z_1$. If $l' = 0$, we choose $l'' = 0$. Thus, let us suppose that $l' \neq 0$. Hence there is an element $\bar{y} \in Z_1$ such that $l'(\bar{y}) < 0$.

Let us show that if $y \in \mathfrak{B}_1$ and $Ly = 0$, then $l'(y) = 0$. Suppose the contrary, so that there is an element $y_1 \in \mathfrak{B}_1$ such that $l'(y_1) = 1$ and $Ly_1 = 0$. Let $y_2 = \bar{y} - 2l'(\bar{y})y_1$. It is evident that $y_2 \in Z_1$ and $l'(y_2) = -l'(\bar{y}) > 0$, contradicting the definition of l' .

If $w \in \mathfrak{B}_2$ and $w = Ly$, let $l''(w) = l'(y)$. By the preceding paragraph, l'' is well defined. It is clear that l'' is defined on all of \mathfrak{B}_2 and is linear. By definition, $l'(y) = l''(Ly)$ for all $y \in \mathfrak{B}_1$. If $w \in Z_2$, then $w = Ly$ for some $y \in Z_1$, so that $l''(w) = l'(y) \leq 0$. Thus it only remains to prove that l'' is continuous. It suffices to show that $\{w: w \in \mathfrak{B}_2, l''(w) < 0\} = H$ is open (see [4, p. 417, Lemma 7]). Now $L^{-1}H = \{y: y \in \mathfrak{B}_1, l'(y) < 0\}$ is open since l' is continuous, and it follows at once from the interior mapping theorem [4, p. 55, Theorem 1] that $L(L^{-1}H) = H$ is open. This completes the proof of the lemma.

Let \mathfrak{B}_2 be the space of all continuous, real-valued functions defined on I_e . Under the sup norm, \mathfrak{B}_2 is a Banach space. Let

$$Z_2 = \{w: w \in \mathfrak{B}_2, \max_{t \in I_e} w(t) < 0\} \cup \{0\}.$$

Clearly, Z_2 is a convex cone in \mathfrak{B}_2 with vertex at 0 and a nonempty interior, and $Z_2 \neq \mathfrak{B}_2$. For every $y \in \mathfrak{C}$, let Ly be the restriction on I_e of the function $\tilde{g}_x(z(t), t) \cdot y(t)$. Clearly, L is a linear, continuous map from \mathfrak{C} into \mathfrak{B}_2 , and, since $\tilde{g}_x(z(t), t) \neq 0$ for all $t \in I_e$ by hypothesis, L is onto \mathfrak{B}_2 .

Appealing to Lemma 8.1, and making use of (8.29), we conclude that there is a functional $l'' \in \mathfrak{B}_2^*$ such that

$$(8.33) \quad -l_{-\mu-1}(y) = l''(Ly) \quad \text{for all } y \in \mathfrak{C},$$

$$(8.34) \quad l''(w) \leq 0 \quad \text{whenever } \max_{t \in I_e} w(t) < 0 \quad \text{and } w \in \mathfrak{B}_2.$$

It is well known [4, p. 265, Theorem 3] that there is a regular, countably additive, scalar-valued measure ν_0 defined on the Borel subsets of I_e such that $l''(w) = \int_{I_e} w(t) d\nu_0$ for all $w \in \mathfrak{B}_2$. Let us define the regular, countably additive measure ν on the Borel subsets of I by means of the relation $\nu(E) = \nu_0(E \cap I_e)$. It follows that if w is any continuous, real-valued function defined on I , and w_e is the restriction of w on I_e , then $l''(w_e) = \int_I w d\nu$, and (8.33) and (8.34) take the form

$$(8.35) \quad l_{-\mu-1}(y) = -\int_I [\tilde{g}_x(z(t), t) \cdot y(t)] \, d\nu \quad \text{for every } y \in \mathcal{C},$$

$$(8.36) \quad \int_I w(t) \, d\nu \leq 0 \quad \text{whenever} \quad \max_{t \in I_e} w(t) < 0,$$

and w is continuous on I .

Let $\lambda(t) = \nu((t, t_2])$ for $t \in (t_1, t_2)$, $\lambda(t_2) = 0$, $\lambda(t_1) = \nu(I)$. It is evident that λ is continuous from the right in (t_1, t_2) and constant on every subinterval of I which does not meet I_e . Since the total variation of ν is bounded [4, p. 128, Lemma 7], the function λ is of bounded variation on I . Further, if w is any continuous real-valued function defined on I , then it is not difficult to show (e.g., by approximating w by step functions and using the Lebesgue dominated convergence theorem) that

$$(8.37) \quad -\int_I w(t) \, d\nu = \int_{t_1}^{t_2} w(t) \, d\lambda(t),$$

where the right-hand integral in (8.37) is in the sense of Riemann-Stieltjes. Thus, we have verified (8.31); (8.32) follows from (8.35) and (8.37). It only remains to verify that λ is nonincreasing. But this is an easy consequence of (8.36) and the definition of λ .

Since $\delta x_{\xi, g^*} \in K^*$ for all $\xi \in R^n$, it follows from (8.28), (8.4), and (8.32), by virtue of (3.2) and (3.3), that

$$\left\{ \sum_{i=\mu}^m \alpha_i [\chi_i' + \chi_i'' \Phi(t_2)] + \int_{t_1}^{t_2} \tilde{g}_x(z(t), t) \Phi(t) \, d\lambda(t) \right\} \xi \leq 0 \quad \text{for all } \xi \in R^n,$$

which is possible only if

$$(8.38) \quad \sum_{i=\mu}^m \alpha_i [\chi_i' + \chi_i'' \Phi(t_2)] + \int_{t_1}^{t_2} \tilde{g}_x(z(t), t) \Phi(t) \, d\lambda(t) = 0.$$

Further, $\delta x_{0, g} \in K^*$ whenever $g \in \Gamma$, so that if we define the functions $\psi(\tau)$ and $\psi_1(\tau, t)$ as before, by (8.7) and (8.8), we obtain (by virtue of (8.28), (8.4), (8.32), (3.2) and (3.3))

$$(8.39) \quad \begin{aligned} & \int_{t_1}^{t_2} \psi(\tau) [g(z(\tau), \tau) - g^*(z(\tau), \tau)] \, d\tau \\ & + \int_{t_1}^{t_2} \int_{t_1}^t \psi_1(\tau, t) [g(z(\tau), \tau) - g^*(z(\tau), \tau)] \, d\tau \, d\lambda(t) \leq 0 \end{aligned}$$

for all $g \in \Gamma$,

as well as relations (8.10)–(8.13). Interchanging the order of integration in the double integral in (8.39), and setting

$$(8.40) \quad \theta(\tau) = \int_{\tau}^{\tau_2} \psi_1(\tau, t) d\lambda(t) \quad \text{for } \tau \in I,$$

$$(8.41) \quad \bar{\psi}(\tau) = \psi(\tau) + \theta(\tau) \quad \text{for } \tau \in I,$$

we obtain

$$(8.42) \quad \int_{t_1}^{\tau_2} \bar{\psi}(\tau) g(z(\tau), \tau) d\tau \leq \int_{t_1}^{\tau_2} \bar{\psi}(\tau) g^*(z(\tau), \tau) d\tau \quad \text{for all } g \in \Gamma.$$

Further (see (8.7), (3.3), (8.38), (8.8) and (8.12))

$$(8.43) \quad \bar{\psi}(t_1) = - \sum_{i=-\mu}^m \alpha_i \chi_i', \quad \bar{\psi}(t_2) = \sum_{i=-\mu}^m \alpha_i \chi_i'',$$

$$\alpha_{-i} \leq 0 \quad \text{for } i \geq 0, \quad \alpha_{-i} = 0 \quad \text{for } i \in \mathcal{J}_z.$$

Let us show that $\bar{\psi}(t) \neq 0$ on a subset of I of positive measure (in the contrary case, (8.42) is trivially satisfied). First consider the case where $\alpha_i = 0$ for $i = -\mu, \dots, m$, so that (see (8.7) and (8.41)) $\psi(\tau) \equiv 0$ and $\bar{\psi}(\tau) \equiv \theta(\tau)$ on I . It then follows from (8.30), on the basis of (8.32), that $\lambda(t) \neq 0$ on I . Also (see (8.38), (8.8) and (8.40)), $\theta(t_1) = \theta(t_2) = 0$. Since λ is nonincreasing on I and $\tilde{g}_x(z(t), t) \neq 0$ for $t \in I_e$, it is now easy to show, on the basis of (8.40) and (8.31), that $\bar{\psi}(\tau) = \theta(\tau) \neq 0$ on a subset of I of positive measure.

Now suppose that $\alpha_i \neq 0$ for some i . It then follows from our hypotheses regarding (8.2), by virtue of (8.43), that either $\bar{\psi}(t_1) \neq 0$ or $\bar{\psi}(t_2) \neq 0$. First consider the case where $t_1 \notin I_e$, $t_2 \notin I_e$. Since I_e is closed, $\theta(t)$ is continuous for $t_1 \leq t \leq t_1 + \epsilon$ and $t_2 - \epsilon \leq t \leq t_2$ and some $\epsilon > 0$ (see (8.31), (8.8), and (8.40)), so that $\bar{\psi}(t)$ is continuous in both of these intervals. Hence $\bar{\psi}$ is different from zero in a neighborhood (in I) either of t_1 or of t_2 . If $t_1 \in I_e$ and $t_2 \in I_e$, it follows from (8.43) and our hypotheses regarding (8.2) that either

$$(8.44) \quad -\bar{\psi}(t_1) + [\alpha' - \lambda(t_1)]\tilde{g}' \neq 0 \quad \text{for every } \alpha' \leq \lambda(t_1)$$

or

$$(8.45) \quad \bar{\psi}(t_2) - \alpha''\tilde{g}'' \neq 0 \quad \text{for every } \alpha'' \geq 0.$$

Let

$$(8.46) \quad \tilde{\psi}(\tau) = \bar{\psi}(\tau) + \lambda(\tau)\tilde{g}_x(z(\tau), \tau) \quad \text{for } \tau \in I.$$

It is not difficult to verify, on the basis of (8.40) and (8.41), that $\tilde{\psi}(\tau)$ is continuous on I . We have just shown that either (see (8.1) and (8.31))

$$\tilde{\psi}(t_1) \neq \alpha'\tilde{g}_x(z(t_1), t_1) \quad \text{for every } \alpha' \leq \lambda(t_1),$$

or

$$\tilde{\psi}(t_2) \neq \alpha'' \tilde{g}_x(z(t_2), t_2) \quad \text{for every } \alpha'' \geq 0.$$

Since $0 \leq \lambda(t) \leq \lambda(t_1)$ for all $t \in I$ and both $\tilde{\psi}(t)$ and $\tilde{g}_x(z(t), t)$ are continuous functions of $t \in I$, we conclude that $\tilde{\psi}(t) = \tilde{\psi}(t) - \lambda(t)\tilde{g}_x(z(t), t) \neq 0$ for t in some neighborhood (in I) either of t_1 or of t_2 . In case $t_1 \in I_e$ and $t_2 \notin I_e$, or $t_1 \notin I_e$ and $t_2 \in I_e$, we can arrive at the same conclusion by arguments virtually unchanged from those presented above.

In summary, we have shown that if z is a solution of Problem 8.1 satisfying the hypotheses indicated above, and if (3.1) holds, then there exist functions $\psi(\tau)$, $\tilde{\psi}(\tau)$, $\psi_1(\tau, t)$ and $\lambda(\tau)$, each defined for all $\tau \in I$, such that

- (a) ψ , $\tilde{\psi}$ and ψ_1 are (row) n -vector valued,
- (b) ψ is an absolutely continuous function satisfying (8.10),
- (c) $\psi_1(\tau, t)$ is an absolutely continuous function of τ for each $t \in I$, and satisfies (8.11) and (8.13),
- (d) $\lambda(\tau)$ is a scalar-valued, nonincreasing function, continuous from the right in (t_1, t_2) , and satisfying (8.31), where I_e is given by (5.11) and (5.9),
- (e) $\tilde{\psi}$ is given by (8.40) and (8.41), satisfies (8.43)—where the α_i are constants which can all vanish only if $\lambda(t_1) \neq 0$ —and is different from zero on a subset of I of positive measure,
- (f) relation (8.42) is satisfied.

Relation (8.42) is a maximum condition; (8.13) and (8.43) (with (8.1)) play the role of transversality conditions.

If the function $\tilde{g}(x, t)$ is of class C^2 with respect to x and t , we can somewhat simplify the form of the necessary conditions. Namely, let us integrate by parts in (8.40), thereby obtaining, by virtue of (8.31), (8.8), (8.41) and (8.46), the relation

$$\tilde{\psi}(\tau) = \psi(\tau) - \int_{\tau}^{t_2} \lambda(t) \frac{\partial \psi_1(\tau, t)}{\partial t} dt.$$

Consequently, $\tilde{\psi}(\tau)$ is absolutely continuous in I , and (see (8.8), (8.10), (3.3), and (3.1))

$$(8.47) \quad \frac{d\tilde{\psi}(t)}{dt} = -\tilde{\psi}(t)g_x^*(z(t), t) + \lambda(t)p_x^*(z(t), t) \quad \text{for almost all } t \in I,$$

where

$$(8.48) \quad p^*(x, t) = \tilde{g}_x(x, t)g^*(x, t) + \tilde{g}_t(x, t).$$

Also, it follows from (8.46), (8.43), (8.1) and (8.31) that

$$(8.49) \quad \begin{aligned} \tilde{\psi}(t_1) &= - \sum_{i=-\mu}^m \alpha_i \chi_i' + \lambda(t_1) \tilde{g}', & \tilde{\psi}(t_2) &= \sum_{i=-\mu}^m \alpha_i \chi_i'', \\ \alpha_{-i} &\leq 0 \quad \text{for } i \geq 0, & \alpha_{-i} &= 0 \quad \text{for } i \in \mathcal{J}_z. \end{aligned}$$

Further, relation (8.42) can be rewritten in the form (see (8.46))

$$(8.50) \quad \begin{aligned} &\int_{t_1}^{t_2} [\tilde{\psi}(t) - \lambda(t) \tilde{g}_x(z(t), t)] g(z(t), t) dt \\ &\quad \leq \int_{t_1}^{t_2} [\tilde{\psi}(t) - \lambda(t) \tilde{g}_x(z(t), t)] g^*(z(t), t) dt \quad \text{for all } g \in \Gamma. \end{aligned}$$

Thus, if z is a solution of Problem 8.1 satisfying the hypotheses previously indicated, if (3.1) holds, and if $\tilde{g}(x, t)$ is of class C^2 with respect to x and t , then there exist an absolutely continuous, (row) n -vector valued function $\tilde{\psi}(t)$ defined on I , and a scalar-valued function $\lambda(t)$ nonincreasing in I , continuous from the right in (t_1, t_2) and satisfying (8.31) (where I_e is given by (5.11) and (5.9)), such that relations (8.47)–(8.50) hold—where the α_i are constants which can all vanish only if $\lambda(t_1) \neq 0$. Also, $\tilde{\psi}(t) \neq \lambda(t) \tilde{g}_x(z(t), t)$ for t in a subset of I of positive measure. Here, (8.50) is the maximum condition; (8.49) (with (8.1)) are the transversality conditions.

If the class Γ in Problem 8.1 is the class Γ^* described in Note 3.1, where the corresponding function $f(x, u, t)$ is continuous in all of its arguments and $U_t = U$ for all $t \in I$, then relation (8.42) (or (8.50), if applicable) can be replaced by a “pointwise” rather than an “integral” maximum condition. Namely, if (7.23) holds, we can show (in the same way that (8.24) was derived from (8.9)) that (8.42) and (8.50) imply the relations

$$(8.51) \quad \tilde{\psi}(t) f(z(t), u^*(t), t) = \max_{v \in U} \tilde{\psi}(t) f(z(t), v, t) \quad \text{for almost all } t \in I,$$

$$(8.52) \quad \begin{aligned} &[\tilde{\psi}(t) - \lambda(t) \tilde{g}_x(z(t), t)] f(z(t), u^*(t), t) \\ &= \max_{v \in U} [\tilde{\psi}(t) - \lambda(t) \tilde{g}_x(z(t), t)] f(z(t), v, t) \quad \text{for almost all } t \in I, \end{aligned}$$

respectively.

Observe that Theorems 4.1 and 4.2 of [1] have given rise to what appear to be two distinct sets of necessary conditions satisfied by solutions of Problem 8.1.

The last set of necessary conditions ((8.52), (8.31), (8.47)–(8.49)) were obtained by Warga [10] under hypotheses which differ from those made by us in the following essential respects. In our notation, Warga assumed that

$$(8.53) \quad \{f(x, v, t) : v \in U\} \text{ is a convex subset of } R^n \text{ for each } (x, t) \in G \times I,$$

and that for each i , either $x_i' = 0$ or $x_i'' = 0$, and $x_0' = 0$. On the other hand, Warga's assumptions as related to (8.2) were somewhat weaker than ours. In all other essential respects, Warga's hypotheses coincided with those made by us in order to obtain (8.52). For certain cases (essentially when his set M is not empty), the necessary conditions he obtained are stronger than ours. However, it appears to us that we can also obtain the stronger necessary conditions by introducing a different topology on \mathcal{C} .

Gamkrelidze, [11] and [2, Chap. VI], obtained necessary conditions for the same problem as the one considered by Warga and without imposing the convexity requirement (8.53). However, Gamkrelidze made a number of alternate, relatively strong hypotheses. Under these additional hypotheses, our necessary conditions imply the basic result of Gamkrelidze [2, Theorem 25, p. 311]. In this regard, it is pertinent to point out that, on a subinterval of I which does not meet I_e , $\bar{\psi}(\tau)$ as given by (8.41) is absolutely continuous and satisfies the equation

$$(8.54) \quad \frac{d\bar{\psi}(\tau)}{d\tau} = -\bar{\psi}(\tau)g_x^*(z(\tau), \tau) \quad \text{for almost all } \tau.$$

Equation (8.54) is easily obtained if one takes into account (8.31).

The same problem was also investigated by Dubovitskii and Milyutin [18] who adopted a viewpoint very similar to the one used here. Indeed, in [18] the "restriction on the phase coordinates" was also formulated in terms of an inequality constraint on a function space by means of the functional defined by (5.9). Relation (5.8) for this functional, where c is given by (5.10), was derived in [18] and was then utilized to obtain necessary conditions. The principal difference between our approach and that of [18] is in the way in which the equality constraints and the differential equation constraints are taken into account. No theorems corresponding to Theorems 4.1–4.6 of [1] were formulated in [18]. Consequently, Dubovitskii and Milyutin were obliged to make some ingenuous, but rather cumbersome constructions (so as to be able to use a fundamental result in [2]) to derive their necessary conditions, which, of course, coincide with ours, i.e., with (8.52), (8.31) together with the fact that λ is nonincreasing and (8.47)–(8.49), with (7.23) (in actuality, in [18], (8.51), (8.40), (8.41), (8.10), (8.8) and (8.43) were given instead of (8.47)–(8.49) and (8.52)).

Some preliminary results for Problem 8.1 along the lines given here were stated (without proof) in [19].

Note 8.1. For the sake of simplicity, we have confined ourselves in this section to a problem with fixed initial and terminal times; problems with free boundary times (as well as with constraints involving intermediate times) can be similarly treated by considering the space $\mathcal{C} \times R^k$ and functionals of the form of (5.6) rather than \mathcal{C} and (5.2), respectively.

Note 8.2. If Problem 8.1 is modified so as to include an additional, finite number of constraints of the form $\varphi_{-\mu-j}(x) \leq 0, j = 2, \dots, \bar{\mu}$, where $\varphi_{-\mu-j}$ has the form of (5.9) for each $j \geq 1$ (and the corresponding functions \tilde{g} satisfy suitable hypotheses), necessary conditions analogous to those obtained for Problem 8.1 can be derived. Problems of this type were considered by Warga [12] under hypotheses similar to those made in [10] (as well as by Gamkrelidze in [2, Chap. VI] under relatively severe hypotheses). The necessary conditions obtained by Warga and Gamkrelidze again essentially follow from the general theorems of [1].

Note 8.3. We can here also make remarks analogous to those made in Note 6.4.

Now consider the following problem, which is commonly referred to as a minimax problem. Let G, I, \mathcal{J} , and \mathcal{C} be defined as before, let Γ be a quasi-convex family in \mathcal{J} , and let $Q^{*'} be correspondingly defined as in §3.$

Problem 8.2. Let $\mathfrak{J} = \mathcal{C}$, and let $\varphi_{-\mu}, \dots, \varphi_{-1}, \varphi_1, \dots, \varphi_m$ be functionals on \mathfrak{J} of the form of (5.2), where the corresponding real-valued functions χ_i are of class C^1 and have a common open domain $G_0 \subset G \times G, k = 2, \tau_1 = t_1$ and $\tau_2 = t_2$. Let φ_0 be a functional on \mathfrak{J} of the form of (5.9), where $\tilde{g}(x, t)$ is a real-valued continuous function defined on $G \times I$ such that \tilde{g}_x is defined and continuous in $G \times I$, and \tilde{I} is a fixed subset of I . Then find an element $x \in W \cap Q^{*'}$ (where $W \subset \mathfrak{J}$ is the common domain of the φ_i) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$ and $\varphi_{-i}(x) \leq 0$ for $i = 1, \dots, \mu$, and such that $\varphi_0(x)$ is minimized.

If we set $Q' = Q^{*'}$, Problem 8.2 falls under the category of the canonical optimization problem described in [1]. Let z be a solution of this problem, so that (3.1) holds. For ease of notation, and without loss of generality, we shall suppose that $\varphi_0(z) = 0$. Let I_e be defined by (5.11) and (5.9). We shall then suppose that $\tilde{g}_x(z(t), t) \neq 0$ for all $t \in I_e$, and that the same assumptions concerning relations (8.2) and (8.1) that were made for Problem 8.1 also hold here (except that the index i here does not assume the value 0). Consequently, z is a totally regular local solution of the canonical optimization problem in the sense of [1].

If we now appeal to Theorems 4.1 and 4.2 of [1], we obtain necessary conditions identical with those obtained for solutions of Problem 8.1, except that now $\alpha_0 = 0$. This, of course, is not unexpected—see [1, Note 4.4].

Previous results for a special case of this problem were obtained by Warga [10], who reduced his problem to one with restrictions on the phase coordinates. This problem was also investigated in [18] and in [21].

The comments of Notes 8.1–8.3 pertain to Problem 8.2 also.

9. Applications to miscellaneous problems. Let us mention briefly some other problems to which the theorems of [1, §4] may be applied.

Let G and I be defined as in §2, let Δ be a fixed positive number, let Γ be a given class of functions $g(x_1, x_2, t)$ from $G \times G \times I$ into R^n , and let $\tilde{\mathcal{C}}$ denote the space of continuous functions $x(t)$ from $[t_1 - \Delta, t_2]$ into R^n , with $\|x\| = \sup_{t_1 - \Delta \leq t \leq t_2} |x(t)|$. Let \tilde{Q}^{**} denote the set of all absolutely continuous functions $x \in \tilde{\mathcal{C}}$ whose range is contained in G , and which satisfy the relation

$$\dot{x}(t) = g(x(t), x(t - \Delta), t) \quad \text{for almost all } t \in I$$

and some function $g \in \Gamma$. If the class Γ is quasiconvex, we can construct a first-order, convex approximation to $\tilde{Q}^{**} - z$ (for any fixed element $z \in \tilde{Q}^{**}$) analogous to the set K^* defined in §3, and obtain on the basis of [1, Theorems 4.1 and 4.2] necessary conditions to problems analogous to Problems 7.1, 8.1, and 8.2. These conditions generalize those obtained by Kharatishvili and described in [2, §27].

We may also consider the following variant of Problem 8.1. Let us suppose that, in the statement of Problem 8.1, the class Γ which determines Q^{**} is Γ^* as described in Note 3.1. Let $p(x, u, t)$ be a real-valued function defined on $G \times U \times I$, and let Q^{***} denote the set of all $x \in Q^{**}$ such that, for some $u \in \Omega^*$, the relations $\dot{x}(t) = f(x(t), u(t), t)$ and $p(x(t), u(t), t) \leq 0$ hold for almost all $t \in I$. Deleting the functional $\varphi_{-\mu-1}$ (given by (5.9)), let us consider the problem of finding an element $x \in W \cap Q^{***}$ (where W is the common domain of $\varphi_{-\mu}, \dots, \varphi_m$, defined as in Problem 8.1) such that $\varphi_i(x) = 0$ for $i = 1, \dots, m$, and $\varphi_{-i}(x) \leq 0$ for $i = 1, \dots, \mu$, and such that $\varphi_0(x)$ is minimized. (We may also modify this problem as indicated in Notes 8.1–8.3.) Such problems have been investigated by, among others, Hestenes [13], Guinn [8], and Gamkrelidze [2, §35].

If the function p is independent of u , this problem reverts to Problem 8.1; if p is independent of x the problem becomes a special case of Problem 7.1.

Under suitable hypotheses (essentially the same as those indicated in [8]) on p, f , and U , we can obtain, on the basis of [1, Theorem 4.2], necessary conditions satisfied by solutions of this problem. These conditions are essentially the same as those given in [8] and [13].

By the same token, we can obtain necessary conditions for the analogous variant of Problem 8.2. Such minimax problems, for particular functions p and f (p was independent of x and t , and f was linear in x and u) were considered, for example, in [14, §II]. A more general problem of this type was investigated in [18].

Finally, we mention a class of optimization problems in which the “phase variable” x may have discontinuities, and the “control” u may include “delta functions”. Certain ones of these problems may be reduced to the more conventional type of problem described in §6 by means of a suitable change of independent variable (this was done by Rishel [15] and Warga [16]), or they may be treated directly by considering functionals which are de-

finied in terms of the total variation of a vector-valued function [17]. For the last-named formulation, the theorems of [1, §4] can again be brought to bear.

REFERENCES

- [1] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I. General theory*, this Journal, 4 (1966), pp. 505-527.
- [2] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [3] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, this Journal, 3 (1965), pp. 106-128.
- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.
- [5] B. W. JORDAN AND E. POLAK, *Optimal control of aperiodic discrete-time systems*, this Journal, 2 (1964), pp. 332-346.
- [6] S. S. L. CHANG, *General theory of optimal processes*, Ibid., 4 (1966), pp. 46-55.
- [7] V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND L. S. PONTRYAGIN, *The theory of optimal processes. I. The maximum principle*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 3-42. (English transl. in Amer. Math. Soc. Transl. (2), 18 (1961), pp. 341-382.)
- [8] T. GUINN, *Weakened hypotheses for the variational problem considered by Hestenes*, this Journal, 3 (1965), pp. 418-423.
- [9] R. V. GAMKRELIDZE, private communication.
- [10] J. WARGA, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432-455.
- [11] R. V. GAMKRELIDZE, *Optimal control processes with restricted phase coordinates*, Izv. Akad. Nauk SSSR Ser. Mat., 24 (1960), pp. 315-356.
- [12] J. WARGA, *Unilateral variational problems with several inequalities*, Michigan Math. J., 12 (1965), pp. 449-480.
- [13] M. R. HESTENES, *On variational theory and optimal control theory*, this Journal, 3 (1965), pp. 23-48.
- [14] L. W. NEUSTADT, *Minimum effort control systems*, Ibid., 1 (1962), pp. 16-31.
- [15] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, Ibid., 3 (1965), pp. 191-205.
- [16] J. WARGA, *Variational problems with unbounded controls*, Ibid., 3 (1965), pp. 424-438.
- [17] L. W. NEUSTADT, *A general theory of minimum-fuel space trajectories*, Ibid., 3 (1965), pp. 317-356.
- [18] A. YA. DUBOVITSKII AND A. A. MILYUTIN, *Extremum problems in the presence of constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395-453.
- [19] L. W. NEUSTADT, *Optimal control problems as extremal problems in a Banach space*, Proceedings of Polytechnic Institute of Brooklyn Symposium on System Theory, 1965, pp. 215-224.
- [20] H. HALKIN, *A maximum principle of the Pontryagin type for systems described by nonlinear difference equations*, this Journal, 4 (1966), pp. 90-111.
- [21] B. N. PSHENICHNIY, *Linear optimal control problems*, Ibid., 4 (1966), pp. 577-593.

MATHEMATICAL THEORY OF OPTIMAL CONTROL FOR SEMILINEAR HYPERBOLIC SYSTEMS IN TWO INDEPENDENT VARIABLES*

WAYNE SCHMAEDEKE†

1. Introduction. Problems involving optimization of systems of partial differential equations have received only a moderate amount of attention up to the present. These studies have as their ultimate goals several aspects of the theory of optimal control. References [1] and [2] consider the uniqueness aspect, [3] and [4] treat necessary conditions, and [5], [6] and [7] mainly consider existence conditions. Reference [7] has the most in common with the present work; in fact, the linear equation along with the boundary conditions treated in [7] is a special case of the almost linear or semilinear equation considered herein.

In this work we shall be concerned with precisely stating conditions which are sufficient for the existence of optimal solutions to mixed initial-boundary value problems for control processes described by semilinear hyperbolic partial differential equations in two independent variables. The cost functional is assumed to be a convex integral and the system of equations is assumed to be in diagonal form. In addition to being given initial conditions on the solution, we assume that boundary relations are given on two arcs Λ_1 and Λ_2 issuing from the endpoints of the "initial interval."

It is important to notice a complicated situation concerning the association between the boundary relations for a given *original* system and the corresponding diagonal system. It is well known that boundary relations cannot be arbitrarily imposed along Λ_1 and Λ_2 for the diagonal system. This brings up the question of whether or not a given set of original boundary relations will result in *admissible* boundary relations after the transformation from the original equations to the diagonal equations has been made. Only partial answers to this question can be found in the literature; for example, the treatment in [8]–[10] should be mentioned. For obvious reasons, we formulate our control problem in the diagonal form, which we shall hereafter refer to as the *canonical* form.

2. Statement of the problem. Let D be an open connected region in the upper half of the (x, t) -plane (i.e., $t \geq 0$) bounded by the segment $x_1 \leq x \leq x_2$ of the x -axis, two smooth arcs Λ_1 and Λ_2 issuing from the end-

* Received by the editors October 24, 1966.

† School of Mathematics, Institute of Technology, University of Minnesota, Minneapolis, Minnesota. This work was supported by the National Aeronautics and Space Administration under Grant 24-005-063.

points of the given segment into the upper half-plane, and the line $t = T$. Consider the semilinear canonical hyperbolic system of dimension n ,

$$(2.1) \quad u_t + \Gamma(x, t)u_x = C(x, t, u) + G(x, t)w(t),$$

where

(a) Γ is an $n \times n$ diagonal matrix having diagonal elements $\gamma_1(x, t), \dots, \gamma_n(x, t)$ which admit continuous first partial derivatives with respect to x and t in \bar{D} (the bar here indicates the closure of D in the usual topology of the plane);

(b) C has continuous first and second partial derivatives with respect to x, t, u_1, \dots, u_n in $\bar{D} \times R^n$;

(c) $G(x, t)$ has continuous first partial derivatives with respect to x and t in \bar{D} ;

(d) $w(t)$ is an m -vector whose components are bounded measurable functions of t on $[0, T]$ and is such that its values lie in a prescribed convex compact nonempty set $\Omega \subset R^m$.

We shall assume that each of the n functions $\gamma_i(x, t)$ is not zero in \bar{D} and we define the *characteristic curves* of (2.1) to be the solutions of the ordinary differential equations

$$(2.2) \quad \frac{dx}{dt} = \gamma_i(x, t), \quad i = 1, \dots, n,$$

lying in \bar{D} . We denote by μ the number of characteristic curves issuing from the left endpoint $(x_1, 0)$ of the initial segment into D . It is thereby assumed that no characteristic curve is tangent to either Λ_1 or Λ_2 . Let the first μ diagonal elements $\gamma_1(x, t), \dots, \gamma_\mu(x, t)$ correspond to those curves entering D from $(x_1, 0)$ and then observe that the remaining $n - \mu$ diagonal elements define characteristic curves which enter D from the right endpoint $(x_2, 0)$ of the initial segment.

The mixed initial-boundary value problem for (2.1) is posed as follows: The solution $u(x, t)$ of (2.1) is required to satisfy the initial condition

$$(2.3) \quad u(x, 0) = \phi(x), \quad x_1 < x < x_2.$$

The boundary conditions consist of μ relations among u_1, \dots, u_n along the boundary curve Λ_1 which may be solved to yield the following equations:

$$(2.4) \quad \begin{aligned} u_1(x, t) &= g_1(u_{\mu+1}(x, t), \dots, u_n(x, t), x, t), \\ &\vdots \\ u_\mu(x, t) &= g_\mu(u_{\mu+1}(x, t), \dots, u_n(x, t), x, t), \end{aligned}$$

which are required to hold when (x, t) lies on Λ_1 . Similar conditions are

given on Λ_2 . The function g_i here is assumed to have continuous partials with respect to all arguments.

The precise sense in which a function $u(x, t)$ is regarded to be a solution of (2.1) will be discussed in the next section. Sufficient conditions will be given which provide for the existence of "a solution" (in the given sense) corresponding to each choice of a function $w(t)$, subject to the restrictions listed in (d) above. Such a function $w(t)$ satisfying (d) will be called an *admissible control function* and the corresponding solution $u(x, t)$ to (2.1) satisfying the mixed initial-boundary conditions (2.3), (2.4) will be called its *response*.

Let there be given the *integral cost functional* $I(w)$ defined by

$$(2.5) \quad I(w) = \int_0^T [g_0(t, u(x_0, t)) + h_0(t, w(t))] dt,$$

where $g_0(t, u)$ and $h_0(t, w)$ are continuous for all t in the interval $[0, T]$ and for all u in R^n and all w in Ω . We shall assume that x_0 is a given constant and that $g_0(t, u)$ and $h_0(t, w)$ are convex functions (for each fixed t) of u and w , respectively, which further satisfy

$$(2.6) \quad g_0(t, u) \geq 0, \quad h^0(t, w) \geq a |w|^p$$

for some constants $a > 0$, $p > 1$.

The optimal control problem consists of selecting an admissible $w^*(t)$ in such a way that it has a response $u^*(x, t)$ with the property that the integral cost functional $I(w^*)$, defined by (2.5) and (2.6), is minimized as compared with all other admissible controls $w(t)$ and their corresponding responses $u(x, t)$, i.e.,

$$(2.7) \quad I(w^*) \leq I(w)$$

for all admissible w .

3. A useful concept of solution. One method of proving the existence of a solution to the mixed-initial boundary value problem (2.1), (2.3), (2.4) is to convert it to a corresponding system of integral equations whose solution can be shown to exist by employing the method of successive approximations. However, in taking this approach, it becomes apparent that the function $u(x, t)$ which satisfies the system of integral equations does not have a continuous partial derivative with respect to t . Therefore, one is led to the problem of defining a concept of solution which is useful in the optimal control problem at hand.

Let us begin with a discussion of the Cauchy problem (2.1), (2.3) in a neighborhood of the initial segment $[x_1, x_2]$. For convenience, let us write (2.1) as

$$(3.1) \quad \frac{\partial u_k}{\partial t} + \gamma_k \frac{\partial u_k}{\partial x} + \tilde{C}_k = 0, \quad k = 1, \dots, n,$$

where $\tilde{C}_k = -C_k(x, t, u) - \{G(x, t) w(t)\}_k$ and γ_k is as defined in (2.1). The initial values are

$$(3.2) \quad u(x, 0) = \phi(x) \quad \text{on} \quad x_1 < x < x_2,$$

where $\phi(x)$ is a prescribed absolutely continuous function. We then transform the Cauchy problem to a system of integral equations by integrating (3.1) along the corresponding characteristic curves

$$(3.3) \quad \frac{dx}{dt} = \gamma_k(x, t), \quad k = 1, \dots, n.$$

It will facilitate our notation if we denote the solutions of the ordinary differential equations (3.2) passing through the point (ξ, τ) of the (x, t) -plane by

$$(3.4) \quad x = X_k(t; \xi, \tau), \quad k = 1, \dots, n.$$

By integrating each equation represented in (3.1) along the arc of the corresponding characteristic through (ξ, τ) which connects that point with the initial segment, using t as a parameter along the curve, and taking into account the fact that

$$\frac{du_k}{dt} = \frac{\partial u_k}{\partial t} + \frac{\partial u_k}{\partial x} \frac{dx}{dt} = -\tilde{C}_k,$$

we are thus able to derive

$$(3.5) \quad \begin{aligned} u_k(\xi, \tau) &= \phi_k(X_k(0; \xi, \tau)) \\ &- \int_0^\tau \tilde{C}_k(X_k(\sigma; \xi, \tau), \sigma, u(X_k(\sigma; \xi, \tau), \sigma)) d\sigma. \end{aligned}$$

The method of successive approximations, applied to (3.5) for the purpose of establishing the existence of a solution $u(\xi, \tau)$, is treated in many sources, for example, in [9], [10], and [11]. We wish to apply this method herein for the purpose of establishing the differentiability properties of the functions $u(x, t)$ which satisfy (3.5). To this end, we define a sequence u_k^r of successive approximations to the solution of (3.5) by means of the formulas

$$(3.6) \quad u_k^0(\xi, \tau) = \phi_k(\xi),$$

$$u_k^{r+1}(\xi, \tau) = \phi_k(X_k(0; \xi, \tau))$$

$$(3.7) \quad - \int_0^\tau \tilde{C}_k(X_k(\sigma; \xi, \tau), \sigma, u^r(X_k(\sigma; \xi, \tau), \sigma)) d\sigma.$$

We shall denote the absolute value of a vector v by $|v| = \max(|v_1|, \dots, |v_n|)$, and then define the norm of a vector with respect to a set S in the (x, t) -plane by

$$\|v(x, t)\| = \text{l.u.b. } |v(x, t)| \quad \text{for } (x, t) \text{ in } S.$$

Now let K denote an upper bound on the absolute values of the functions \tilde{C}_k and γ_k in some neighborhood of the segment $[x_1, x_2]$ in the plane. Let K also denote an upper bound on the absolute values of the first partial derivatives of \tilde{C}_k and γ_k with respect to x , on the absolute values of the first partial derivatives of the γ_k with respect to t , and on the absolute values of ϕ_k , as well as the absolute values of the derivatives of ϕ_k with respect to x where these derivatives exist. Finally, let K also denote an upper bound on the absolute values of the first partial derivatives of \tilde{C}_k with respect to u_1, \dots, u_n . These assumptions will all hold in a hexagonal region R about the initial segment defined by the lines $\pm Kt = x - x_1$, $\pm Kt = -x + x_2$, $t = \pm \alpha$ for α sufficiently small.

It is well known that the sequence u_k^r converges in this region and that a continuous solution u of (3.1) exists (cf. Garabedian [11]). The situation we face here however is different from the one in [11] in two respects, namely, the initial functions $\phi_k(x)$ are merely absolutely continuous (A.C.), and the functions \tilde{C}_k are merely measurable in t . We shall thus only be able to show that the continuous functions $u_k(x, t)$ which exist and satisfy (3.1), (3.2) are A.C. in x and t separately and have partial derivatives with respect to x and t which exist almost everywhere (a.e.) but are not necessarily continuous.

To begin our investigation, let us show that the functions $u_k^{r+1}(\xi, \tau)$ are A.C. in ξ and τ separately. We shall need a simple lemma concerning composite functions of A.C. functions (cf. [12]).

LEMMA 1. *Let f be real-valued and A.C. on an interval $[\alpha, \beta]$ and let $g(x)$ be defined on $[a, b]$, be A.C. therein, and have values in $[\alpha, \beta]$. Define the composite function h by*

$$h(x) = f(g(x)), \quad x \text{ in } [a, b].$$

Then if g is monotone or f is Lipschitzian, h is also A.C. on $[a, b]$.

We next recall that the functions $\gamma_k(x, t)$ have continuous first partials with respect to x and t , and thus the solutions $X_k(t; \xi, \tau)$ of (3.3) have continuous first partials with respect to ξ and τ (this is a well-known result in ordinary differential equations). In particular, we observe that $X_k(0; \xi, \tau)$ is A.C. in ξ for each fixed τ and A.C. in τ for each fixed ξ . Moreover, $X_k(0; \xi, \tau)$ is monotone in each variable ξ and τ separately since $X_k(0; \xi, \tau)$ is merely the x -coordinate of the intersection of the k th characteristic curve through (ξ, τ) with the initial segment. By applying Lemma

1 to the composite function $\phi_k(X_k(0; \xi, \tau))$ we obtain the result that ϕ_k is A.C. in ξ for each fixed τ and A.C. in τ for each fixed ξ .

The term represented by the integral in (3.7) is easily shown to be A.C. in ξ and τ separately since $\tilde{C}_k(x, t, u)$ has continuous first partials (as well as second partials) with respect to x and u_1, \dots, u_n and is, in particular, Lipschitz with regard to these arguments. Therefore, the functions $u_k^{r+1}(\xi, \tau)$ are A.C. in ξ and τ separately, and it is a trivial matter to show that the limit function $u_k(\xi, \tau)$ is A.C. in ξ and τ separately.

The next step in our analysis will be concerned with the sequences of partial derivatives $\partial u_k^{r+1}/\partial \xi$ and $\partial u_k^{r+1}/\partial \tau$. We observe that these partials exist and it is only a matter of showing they converge to $\partial u_k/\partial \xi$ and $\partial u_k/\partial \tau$, respectively. Actually, the convergence is only a.e. and can easily be established by virtue of the following lemma.

LEMMA 2. Suppose that the functions $u^r(x, t)$ are A.C. in x on $(0, 1)$ for each fixed t and that the functions $|\partial u^r/\partial x|$ are uniformly bounded by a fixed integrable function $\alpha(x)$. Let the limits

$$\begin{aligned}\lim_{r \rightarrow \infty} u^r(x, t) &= u(x, t) \text{ a.e.,} \\ \lim_{r \rightarrow \infty} \frac{\partial u^r}{\partial x}(x, t) &= v(x, t) \text{ a.e.}\end{aligned}$$

exist for each fixed t . Then $u(x, t)$ is A.C. on $(0, 1)$ and $\partial u/\partial x = v(x, t)$ a.e.

Proof. Since $u^r(x, t)$ are A.C. in x , we may write

$$u^r(x, t) = u^r(0, t) + \int_0^x u_x^r(\xi, t) d\xi \quad \text{for } 0 \leq x \leq 1.$$

Letting $r \rightarrow \infty$ we obtain for almost all x that

$$u(x, t) = u(0, t) + \lim_{r \rightarrow \infty} \int_0^x u_x^r(\xi, t) d\xi.$$

Now, because $u_x^r(x, t) \rightarrow v(x, t)$ a.e. and because $|u_x^r| \leq \alpha(x)$ which is integrable, then by Lebesgue's dominated convergence theorem, $\lim u_x^r(\xi, t)$ is integrable and

$$u(x, t) = u(0, t) + \int_0^x v(\xi, t) d\xi.$$

It follows that $u(x, t)$ is A.C. in x for each fixed t and moreover, being differentiable for almost all x in $[0, 1]$, we have

$$u_x(x, t) = v(x, t) \text{ a.e.}$$

To show that $u_k(\xi, \tau)$ has a first partial derivative with respect to ξ , we

form the first partials of u_k^{r+1} by differentiating (3.7) and taking account of the fact that $\tilde{C}_k = -C_k - \{Gw\}_k$. Thus

$$(3.8) \quad \begin{aligned} \frac{\partial u_k^{r+1}(\xi, \tau)}{\partial \xi} &= \frac{\partial \phi_k}{\partial x}(X_k(0; \xi, \tau)) \frac{\partial X_k}{\partial \xi}(0; \xi, \tau) \\ &+ \int_0^\tau \left[\frac{\partial C_k}{\partial x} \frac{\partial X_k}{\partial \xi} + \sum_{i=1}^n \frac{\partial C_k}{\partial u_i} \frac{\partial u_i^r}{\partial x} \frac{\partial X_k}{\partial \xi} \right] d\sigma \\ &+ \int_0^\tau \left\{ \frac{\partial G}{\partial x} \frac{\partial X_k}{\partial \xi} w(\sigma) \right\}_k d\sigma. \end{aligned}$$

In (3.8) we have reverted to the original notation involving C and G rather than \tilde{C} . We shall therefore restate our boundedness and Lipschitz assumptions, which are based upon the hypotheses in §2. Let K be an upper bound in R of the absolute values of $\{G(x, t)w(t)\}_k$, $\partial X_k/\partial \xi$, C_k , $\partial C_k/\partial u_i$, $\partial u_i^r/\partial x$, and ϕ_k as well as a Lipschitz constant for $\partial C_k/\partial u_i$ and $\partial C_k/\partial x$ in all their arguments. Now consider the difference

$$(3.9) \quad \begin{aligned} \frac{\partial u_k^{r+1}}{\partial \xi} - \frac{\partial u_k^r}{\partial \xi} &= \int_0^\tau \left[\frac{\partial C_k}{\partial x}(X_k, \sigma, u^r) - \frac{\partial C_k}{\partial x}(X_k, \sigma, u^{r-1}) \right] \frac{\partial X_k}{\partial \xi} d\sigma \\ &+ \int_0^\tau \sum_{i=1}^n \frac{\partial C_k}{\partial u_i}(X_k, \sigma, u^r) \left[\frac{\partial u_i^r}{\partial x} - \frac{\partial u_i^{r-1}}{\partial x} \right] \frac{\partial X_k}{\partial \xi} d\sigma \\ &+ \int_0^\tau \sum_{i=1}^n \left[\frac{\partial C_k}{\partial u_i}(X_k, \sigma, u^r) - \frac{\partial C_k}{\partial u_i}(X_k, \sigma, u^{r-1}) \right] \\ &\quad \cdot \frac{\partial u_i^{r-1}}{\partial x} \frac{\partial X_k}{\partial \xi} d\sigma. \end{aligned}$$

By using the definition of the norm of a vector in R we may obtain from (3.9) the inequality

$$(3.10) \quad \begin{aligned} \left\| \frac{\partial u^{r+1}}{\partial \xi} - \frac{\partial u^r}{\partial \xi} \right\| &\leq (\alpha K^2 + \alpha n K^3) \|u^r - u^{r-1}\| \\ &+ \alpha n K^2 \left\| \frac{\partial u^r}{\partial \xi} - \frac{\partial u^{r-1}}{\partial \xi} \right\|. \end{aligned}$$

Moreover, by referring to (3.7) and forming the difference $u_k^{r+1} - u_k^r$, we obtain the inequality

$$|u_k^{r+1} - u_k^r| \leq \int_0^\tau |C_k(X_k, \sigma, u^r) - C_k(X_k, \sigma, u^{r-1})| d\sigma,$$

or, in terms of the norm,

$$(3.11) \quad \|u^{r+1} - u^r\| \leq \alpha K \|u^r - u^{r-1}\|.$$

From (3.6) and (3.7) we obtain the bound

$$\|u^1 - u^0\| \leq 2K + 2\alpha K.$$

We can therefore prove by induction that

$$(3.12) \quad \|u^r - u^{r-1}\| \leq 2(\alpha K)^{r-1}(1 + \alpha)K.$$

Now choose α small enough to make $\alpha K < \frac{1}{2}$ and $\alpha n K^2 < \frac{1}{4}$. Then define $H = \alpha K^2 + \alpha n K^3$, $N = 2(1 + \alpha)K$, and $\delta_{r+1} = \|\partial u^{r+1}/\partial \xi - \partial u^r/\partial \xi\|$. Hence, (3.10) becomes

$$(3.13) \quad \delta_{r+1} \leq \frac{1}{4}\delta_r + H \|u^r - u^{r-1}\|.$$

Furthermore, (3.12) reduces to

$$(3.14) \quad \|u^r - u^{r-1}\| \leq \left(\frac{1}{2}\right)^{r+1} 4N.$$

We can choose a number L so that

$$L \geq \max(4\delta_2, 8HN),$$

and then assert that

$$(3.15) \quad \delta_r \leq \left(\frac{1}{2}\right)^r L \quad \text{for } r > 2.$$

The proof is by induction. When $r = 3$ we have from (3.13) and (3.14)

$$\begin{aligned} \delta_3 &\leq \frac{1}{4}\delta_2 + H \left(\frac{1}{2}\right)^3 4N \\ &\leq \frac{4\delta_2}{2^4} + \frac{8HN}{2^4} \leq \frac{L}{2^3}, \end{aligned}$$

and thus the assertion is true for $r = 3$. Suppose then that (3.15) is true when $r = q > 2$. This yields the following calculation:

$$\begin{aligned} \delta_{q+1} &\leq \frac{1}{4}\delta_q + H \left(\frac{1}{2}\right)^{q+1} 4N \\ &\leq \left(\frac{1}{2}\right)^{q+1} \frac{L}{2} + \left(\frac{1}{2}\right)^{q+1} \frac{L}{2} = \left(\frac{1}{2}\right)^{q+1} L. \end{aligned}$$

This establishes the assertion which in turn implies that the terms of the series $\sum_{p=3}^{\infty} (\partial u^p/\partial \xi - \partial u^{p-1}/\partial \xi)$ are bounded by the terms of a convergent geometric series. Since the r th partial sum of the bounded series is $\partial u^r/\partial \xi - \partial u^2/\partial \xi$, we may conclude that $\partial u^r/\partial \xi$ converges on each line $t = c \leq \alpha$ in so far as this line is contained in R . The application of Lemma 2 next yields the result that $\partial u^r/\partial \xi$ converges to $\partial u/\partial \xi$ a.e. on each line $t = c$ contained in R .

The same type of argument as the preceding may be used to establish that $\partial u / \partial \tau$ exists a.e. on each vertical line $x = c$ contained in R .

We shall summarize the preceding investigation in the following way. A unique solution $u(x, t)$ to the initial value problem (2.1), (2.3) exists in the region R , where by solution we mean a function $u(x, t)$ which is A.C. in x and t separately in R , which has partial derivatives with respect to x and t which exist a.e. in R , and which together with u_t and u_x satisfies (2.1) a.e. in R .

Let us now turn to the mixed initial-boundary value problem (2.1), (2.3), (2.4). We follow the treatment given by Lax in [10]. The existence and uniqueness for a solution may be established in exactly the same way; therefore we shall merely indicate the technique and refer the reader, who desires to see the proofs of some of the assertions made in our discussion, to [10].

Consider the region D described in §2. It is possible to determine the boundary values $u_k(x, t)$ for $(x, t) \equiv P$ on Λ_1 and $k = \mu + 1, \dots, n$ by the formula

$$(3.16) \quad \begin{aligned} u_k(x, t) = & \psi_k(P_k) + \int_{t_k}^t C_k(X_k(\sigma; x, t), \sigma, u(X_k(\sigma; x, t), \sigma)) d\sigma \\ & + \int_{t_k}^t \{G(X_k(\sigma; x, t), \sigma)w(\sigma)\}_k d\sigma, \end{aligned}$$

where $P_k = (X_k(t_k; x, t), t_k)$ is the intersection of the k th characteristic curve through (x, t) (which is on Λ_1) with either Λ_2 or the initial segment. The notation ${}_k\psi$ means ϕ_k if P_k is on the initial segment; otherwise, it is the value $u_k(P_k)$ of u_k on Λ_2 which we may assume is known (for it may be determined by this method in a finite number of steps). A similar formula holds for (x, t) on Λ_2 .

After the boundary values for $u_{\mu+1}, \dots, u_n$ have been obtained on Λ_1 , the values of u_1, \dots, u_μ are obtained from (2.4) as

$$u_1(x, t) = g_1(u_{\mu+1}(x, t), \dots, u_n(x, t), x, t), \text{ etc.}$$

By supposing then that all of the boundary values of u are available, we may represent the solution at an interior point (x, t) of D as

$$(3.17) \quad \begin{aligned} u_k(x, t) = & \psi_k(P_k) + \int_{t_k}^t C_k(X_k(\sigma; x, t), \sigma, u(X_k(\sigma; x, t), \sigma)) d\sigma \\ & + \int_{t_k}^t \{G(X_k(\sigma; x, t), \sigma)w(\sigma)\}_k d\sigma, \end{aligned}$$

where, as before, ψ_k is either ϕ_k (if P_k is on the initial segment) or the value of u_k on the boundary curve Λ_1 (if $k \leq \mu$) or Λ_2 (if $k > \mu$).

4. Existence of an optimal control. In this section we prove that a solution to the optimization problem posed in §2 exists under somewhat restrictive hypotheses. We consider the region \bar{D} defined at the beginning of §2, and we denote by Δ the set of all functions $w(t)$ which are admissible control functions such that their responses exist everywhere in \bar{D} and satisfy the mixed initial-boundary value problem (2.1), (2.3), (2.4). Our assumptions are as follows:

- (a) Δ is not empty (in fact, Δ contains infinitely many elements);
- (b) there exists a real bound $B < \infty$ such that $\|u(x, t)\| \leq B$ for all responses to controls in Δ (the norm being taken with respect to \bar{D});
- (c) there exist two real constants $\theta_1 > 0 > \theta_2$ such that $\gamma_i(x, t) \geq \theta_1$ everywhere in \bar{D} for $i = 1, 2, \dots, \mu$ and also $\theta_2 \geq \gamma_j(x, t)$ everywhere in \bar{D} for $j = \mu + 1, \dots, n$.

Under these conditions we have

$$\inf I(w) = \tilde{m} > -\infty,$$

where the infimum is taken over all controls in Δ . We may then select a sequence of controls $w^r(t)$ for which $I(w^r)$ decreases monotonically to \tilde{m} . Since each $w^r(t)$ is bounded and measurable on $[0, T]$, and thus belongs to a closed ball in the Hilbert space $L_2[0, T]$, we may select a subsequence (still to be labeled w^r) such that $w^r(t) \rightarrow w^*(t)$ weakly in $L_2[0, T]$. It is a simple matter to show that $w^*(t)$ is in Δ (cf. [13] where it is shown that $w^*(t)$ may have to be altered on a set of measure zero) and we shall not give the proof here.

We next consider the response $u^r(x, t)$ to the control $w^r(t)$, which, by virtue of (3.17), we may write as

$$(4.1) \quad \begin{aligned} u_k^r(x, t) = & \psi_k^r(P_k) + \int_{t_k}^t C_k(X_k(\sigma; x, t), \sigma, u^r(X_k(\sigma; x, t), \sigma)) d\sigma \\ & + \int_{t_k}^t \{G(X_k(\sigma; x, t), \sigma)w^r(\sigma)\}_k d\sigma, \quad k = 1, \dots, n. \end{aligned}$$

Recall here that $P_k = (X_k(t_k; x, t), t_k)$ is the intersection of the k th characteristic curve through (x, t) with the boundary of D . We will show that the family $u_k^r(x, t)$ of real-valued functions defined in \bar{D} is equicontinuous. We begin by showing that the functions $u_k^r(x, t)$ are equicontinuous on the boundaries Λ_1, Λ_2 , and the initial segment. Equicontinuity on the initial segment follows from the fact that each $u_k^r(x, t)$ is equal to $\phi_k(x)$ when $t = 0$, and the fact that the function $\phi_k(x)$ is continuous for $x_1 \leq x \leq x_2$. Let us next divide each of the arcs Λ_1 and Λ_2 into a finite number of subarcs in the following way. Denote the point where the straight line of slope θ_1 (see hypothesis (c)) through the point $(x_1, 0)$ intercepts

Λ_2 by Y_1 ; similarly, denote the point on Λ_1 where the straight line through the point $(x_2, 0)$ of slope θ_2 intercepts it by Z_1 . Then construct lines through the points Y_1 and Z_1 of slopes θ_2 and θ_1 respectively and denote the points where these lines intercept Λ_1 and Λ_2 by Z_2 and Y_2 respectively. Continue this procedure until points Z_α, Y_β are obtained whose t -coordinates are greater than or equal to T .

Now observe that every point on Λ_2 between $(x_2, 0)$ and Y_1 has the property that the k th characteristic curve (for $1 \leq k \leq \mu$) through it will intersect the initial segment. Likewise, every point on Λ_1 between $(x_1, 0)$ and Z_1 has the property that the k th characteristic curve through it (for $\mu + 1 \leq k \leq n$) will intersect the initial segment. Hence, if we restrict our attention to points on Λ_1 below Z_1 we can show that the $u_k^r(x, t)$ are equicontinuous for $\mu + 1 \leq k \leq n$. To see this, let P be an arbitrary point below Z_1 and let Q be another point below Z_1 (both P and Q are assumed to be on Λ_1 of course) and form the difference

$$\begin{aligned} u_k^r(P) - u_k^r(Q) &= \phi_k(P_k) - \phi_k(Q_k) \\ &+ \int_0^t [C_k(X_k(\sigma; P), \sigma, u^r(X_k(\sigma; P), \sigma)) - C_k(X_k(\sigma; Q), \sigma, u^r(X_k(\sigma; Q), \sigma))] d\sigma \\ &+ \int_t^\tau C_k(X_k(\sigma; Q), \sigma, u^r(X_k(\sigma; Q), \sigma)) d\sigma \\ &+ \int_0^t \{G(X_k(\sigma; P), \sigma)w^r(\sigma) - G(X_k(\sigma; Q), \sigma)w^r(\sigma)\}_k d\sigma \\ &+ \int_t^\tau \{G(X_k(\sigma; Q), \sigma)w^r(\sigma)\}_k d\sigma, \end{aligned}$$

where we have used $P = (x, t)$, $Q = (z, \tau)$, and P_k and Q_k are the points where the k th characteristics through P and Q respectively intersect the initial segment. Because $C_k(x, t, u)$, $G_{kj}(x, t)$ are Lipschitz in x (here $j = 1, \dots, m$), the preceding difference reduces to

$$\begin{aligned} |u_k^r(P) - u_k^r(Q)| &\leq |\phi_k(P_k) - \phi_k(Q_k)| \\ &+ \int_0^t K |X_k(\sigma; P) - X_k(\sigma; Q)| d\sigma \\ &+ \int_0^t WmK |X_k(\sigma; P) - X_k(\sigma; Q)| d\sigma \\ &+ 2K |\tau - t| \end{aligned}$$

for appropriately defined K (i.e., sufficiently large to serve as the Lipschitz constants and the various upper bounds). Here we have used W as a bound on the components of w^r and we notice that W is finite and independent of

r since w^r belongs to Ω which is compact. Finally, we may use the fact that $X_k(\sigma, P)$ is Lipschitz in P to obtain

$$\begin{aligned} |u_k^r(P) - u_k^r(Q)| &\leq |\phi_k(P_k) - \phi_k(Q_k)| + tK^2 |P - Q| \\ &\quad + tWmK^2 |P - Q| + 2K |\tau - t|. \end{aligned}$$

It is clear that given any $\epsilon > 0$ we may choose a neighborhood N_ϵ of P so that all points Q on Λ_1 which are also in N_ϵ yield the result that τ is sufficiently close to t , Q_k is sufficiently close to P_k , and Q is sufficiently close to P , so that

$$|u_k^r(P) - u_k^r(Q)| \leq \epsilon$$

independent of r . Thus the values of $u_k^r(x, t)$ for $\mu + 1 \leq k \leq n$ on this lower segment of Λ_1 are equicontinuous. A similar analysis applies to the lower segment of Λ_2 between $(x_2, 0)$ and Y_1 .

The next step in our proof is to show that the functions $u_k^r(x, t)$ for $1 \leq k \leq \mu$ are equicontinuous on the part of Λ_1 between $(x_1, 0)$ and Z_1 . From (2.4) we may write for $1 \leq k \leq \mu$

$$u_k^r(P) = g_k(u_{\mu+1}^r(P), \dots, u_n^r(P), P),$$

and since g_k has continuous partials with respect to each of its arguments, we see that equicontinuity of the functions u_j^r for $\mu + 1 \leq j \leq n$ implies equicontinuity of the functions u_k^r for $1 \leq k \leq \mu$. Thus the whole vector valued family $u^r(P)$ is equicontinuous on the part of Λ_1 between $(x_1, 0)$ and Z_1 and on Λ_2 between $(x_2, 0)$ and Y_1 .

The task of showing the equicontinuity of $u^r(P)$ on Λ_1 between Z_1 and Z_2 is easily accomplished once we observe that the k th characteristic through a point P between Z_1 and Z_2 (on Λ_1 and with $\mu + 1 \leq k \leq n$) will intersect the boundary of D either on the initial segment or on the part of Λ_2 between $(x_2, 0)$ and Y_1 . This fact coupled with (4.1) shows that $\psi_k^r(P_k)$ is an element of an equicontinuous family, and thus, mutatis mutandis, the preceding proof establishing equicontinuity on Λ_1 between $(x_1, 0)$ and Z_1 applies.

We may proceed in this manner to show that u^r is an equicontinuous family on the initial segment and on Λ_1 and Λ_2 up to and including the points of intersection with the line $t = T$. Finally, we establish the equicontinuity at interior points of D and along the boundary $t = T$ in exactly the same way since the boundary values $\psi_k^r(P_k)$ corresponding to an interior point P are equicontinuous.

Returning now to the sequence $w^r(t)$ which converges weakly to $w^*(t)$ in $L_2[0, T]$, we may apply Ascoli's theorem to the corresponding sequence $u^r(x, t)$ of responses in \bar{D} to obtain a uniformly convergent subsequence (still to be labeled $u^r(x, t)$) and an associated limit function which we

shall denote by $u^*(x, t)$. We now show that $u^*(x, t)$ is the response to the control $w^*(t)$ on $[0, T]$. From (4.1) we have

$$(4.2) \quad \begin{aligned} u_k^*(x, t) &= \lim_{r \rightarrow \infty} \psi_k^r(P_k) \\ &+ \lim_{r \rightarrow \infty} \int_{t_k}^t C_k(X_k(\sigma; x, t), \sigma, u^r(X_k(\sigma; x, t), \sigma)) d\sigma \\ &+ \lim_{r \rightarrow \infty} \int_{t_k}^t \{G(X_k(\sigma; x, t), \sigma) w^r(\sigma)\}_k d\sigma, \end{aligned}$$

which, by weak convergence of $w^r(\sigma)$ and by Lebesgue's dominated convergence theorem, we may write as

$$(4.3) \quad \begin{aligned} u_k^*(x, t) &= \lim_{r \rightarrow \infty} \psi_k^r(P_k) \\ &+ \int_{t_k}^t C_k(X_k(\sigma; x, t), \sigma, u^*(X_k(\sigma; x, t), \sigma)) d\sigma \\ &+ \int_{t_k}^t \{G(X_k(\sigma; x, t), \sigma) w^*(\sigma)\}_k d\sigma. \end{aligned}$$

In order to compute the limit of the boundary functions $\psi_k^r(P_k)$, we proceed in a manner much like that of the previous investigation of equicontinuity. If P belongs to Λ_1 between $(x_1, 0)$ and Z_1 , then for any k such that $\mu + 1 \leq k \leq n$, the k th characteristic curve through P will intersect the initial segment at a point we will call P_k^1 . Hence the boundary value at P can be written as

$$\begin{aligned} \psi_k^r(P) &= \phi_k(P_k^1) + \int_0^{t_k} C_k(X_k(\sigma; P), \sigma, u^r(X_k(\sigma; P), \sigma)) d\sigma \\ &+ \int_0^{t_k} \{G(X_k(\sigma; P), \sigma) w^r(\sigma)\}_k d\sigma. \end{aligned}$$

Thus, the limiting value in this instance is easily seen to be

$$(4.4) \quad \begin{aligned} \lim_{r \rightarrow \infty} \psi_k^r(P) &= \phi_k(P_k^1) + \int_0^{t_k} C_k(X_k(\sigma; P), \sigma, u^*(X_k(\sigma; P), \sigma)) d\sigma \\ &+ \int_0^{t_k} \{G(X_k(\sigma; P), \sigma) w^*(\sigma)\}_k d\sigma. \end{aligned}$$

By uniqueness of the solution to our partial differential equation, the right-hand side in (4.4) is the boundary value of the function u_k^* at the point P . Now, if $1 \leq k \leq \mu$, we merely use (2.4) to compute

$$(4.5) \quad \begin{aligned} \lim_{r \rightarrow \infty} \psi_k^r(P) &= \lim_{r \rightarrow \infty} g_k(u_{\mu+1}^r(P), \dots, u_n^r(P), P) \\ &= g_k(u_{\mu+1}^*(P), \dots, u_n^*(P), P). \end{aligned}$$

But the right-hand side of (4.5) is the required value of the function u_k^* at the boundary point P . Hence, for $1 \leq k \leq n$, the limiting value of the function $\psi_k^r(P)$ is the appropriate boundary value of the function u_k^* .

By assuming that P is on Δ_2 between $(x_2, 0)$ and Y_1 , we obtain the same result. Then we compute the limit of $\psi_k^r(P)$ for P on Δ_1 between Z_1 and Z_2 and observe that the k th characteristic curve through P for $\mu + 1 \leq k \leq n$ will intersect the boundary of D either along Δ_2 between $(x_2, 0)$ and Y_1 or along the initial segment. In either case, the representation of the boundary value $\psi_k^r(P)$, using (4.1), converges to the value $u_k^*(P)$ of the function u_k^* . Also, exactly as before, when $1 \leq k \leq \mu$, equations (2.4) yield the value of u_k^* at the point P . Hence, we may conclude that at every boundary point on either Δ_1 or Δ_2 , the sequence $\psi_k^r(P)$ converges to the appropriate value required of the function u_k^* . This fact can now be used in (4.3) to assert that $u^*(x, t)$ is the response to the control $w^*(t)$ everywhere in \bar{D} .

Our final task is to compute the cost of $w^*(t)$. From (2.5) we may represent the cost of each w^r as

$$(4.6) \quad I(w^r) = \int_0^T [g_0(t, u^r(x_0, t)) + h_0(t, w^r(t))] dt.$$

It is a property of convex functions that if $w^r(t)$ converges weakly to $w^*(t)$ then

$$(4.7) \quad \liminf_{r \rightarrow \infty} \int_0^T h_0(t, w^r(t)) dt \geq \int_0^T h_0(t, w^*(t)) dt$$

(cf. [14] where the details of the proof are given). Hence, taking the limit in (4.6) (and noting the pointwise convergence of $u^r(x_0, t)$ to $u^*(x_0, t)$), we obtain

$$\begin{aligned} \tilde{m} = \lim_{r \rightarrow \infty} I(w^r) &\geq \lim_{r \rightarrow \infty} \int_0^T g_0(t, u^r(x_0, t)) dt + \liminf_{r \rightarrow \infty} \int_0^T h_0(t, w^r(t)) dt \\ &\geq \int_0^T g_0(t, u^*(x_0, t)) dt + \int_0^T h_0(t, w^*(t)) dt = I(w^*). \end{aligned}$$

Thus, we conclude that $I(w^*) = \tilde{m}$ and that $w^*(t)$ is an optimal control.

REFERENCES

- [1] A. FRIEDMAN, *Optimal control in Banach spaces*, to appear.
- [2] YU. V. EGOROV, *Necessary conditions for optimal control in Banach space*, Mat. Sb., 64(106) (1964), pp. 79-101.
- [3] K. A. LUR'E, *The Mayer-Bolza problem for multiple integrals and the optimization of the performance of systems with distributed parameters*, Prikl. Mat. Meh., 27 (1963), pp. 842-853.
- [4] A. I. EGOROV, *On optimal control of processes in distributed objects*, Ibid., 27 (1963), pp. 688-696; J. Appl. Math. Mech., 27 (1964), pp. 1045-1058.

- [5] J. L. LIONS, *Sur quelques problèmes d'optimisation dans les équations d'évolution linéaires de type parabolique*, Functional Analysis and Optimization, E. R. Caianiello, ed., Academic Press, New York, 1965.
- [6] ———, *Optimisation pour certaines classes d'équations d'évolution non linéaires*, to appear.
- [7] D. L. RUSSELL, *Optimal regulation of linear symmetric hyperbolic systems with finite dimensional controls*, Technical Summary Report 566, Mathematics Research Center, University of Wisconsin, Madison, 1965.
- [8] L. L. CAMPBELL AND A. ROBINSON, *Mixed problems for hyperbolic partial differential equations*, Proc. London Math. Soc. (3), 5 (1955), pp. 129–147.
- [9] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, vol. II, Interscience, New York, 1962.
- [10] P. D. LAX, *Partial Differential Equations*, Institute of Mathematical Sciences, New York University, New York, 1953.
- [11] P. R. GARABEDIAN, *Partial Differential Equations*, John Wiley, New York, 1964.
- [12] E. J. MCSHANE AND T. A. BOTTS, *Real Analysis*, D. Van Nostrand, Princeton, 1959.
- [13] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.
- [14] ———, *Foundations of Optimum Control Theory*, John Wiley, New York, 1966.

ON THE EXISTENCE OF SOLUTIONS TO A DIFFERENTIAL GAME*

P. P. VARAIYA†

Abstract. In this paper we consider the problem of the existence of a “min-sup” strategy to a pursuit-evasion game. The dynamics of the players have been modeled by a general dynamical system rather than by a differential system. This has helped to achieve mathematical simplicity as well as clarification of the problems involved in a competitive situation. We have discussed the relation between the two models and the relevance of our results to time-optimal control problems.

1. Introduction. In this paper we study the problem of the existence of a solution to a pursuit-evasion game. The rules of the game can be framed as follows:

There are two players, one called the *pursuer* and the other called the *evader*. The states of these players at any time t , $0 \leq t < \infty$, are represented by n -dimensional vectors $p(t)$ and $e(t)$, respectively. The game starts at time $t = 0$. The dynamics of the two players are given by certain axioms. These axioms are a weaker version of those given in [1] but more restrictive than those given in [2]. The basic notion used in the formulation of these axioms is that of the *attainability function* $P(p_0, t_0, t)$ (or $E(e_0, t_0, t)$) which represents the set of states that can be reached at time t by the pursuer (or evader) starting in state p_0 (or e_0) at time t_0 . A *motion* for the pursuer (or evader) is therefore a mapping $u(\cdot)$ (or $v(\cdot)$) of an interval of $[0, \infty)$ into R^n such that

$$u(t) \in P(u(t_0), t_0, t), \quad t \geq t_0,$$

(or

$$v(t) \in E(v(t_0), t_0, t), \quad t \geq t_0).$$

The evader is informed of the dynamics of the pursuer and the initial state of the pursuer (as well as his own dynamics, of course). This is the extent of the evader's knowledge. A strategy for the evader therefore, consists in selecting, a priori, a motion which satisfies his constraints. The pursuer, on the other hand, together with being supplied with the dynamics of both players, is also told at each instant of time the motion of the evader up to that time. Based on this knowledge, the pursuer selects a course of action which takes him within a specified region (called the endzone) of

* Received by the editors June 21, 1966.

† Electronics Research Laboratory, College of Engineering, University of California, Berkeley, California 94720. This research was wholly supported by the Joint Services Electronics Program (U. S. Army, U. S. Navy and U. S. Air Force) under Grant AF-AFOSR-139-66.

the evader in the shortest possible time. The evader, of course, tries to escape from this predicament as long as possible. The game ends as soon as the pursuer has achieved his goal. For each strategy (= motion) v of the evader and each strategy g of the pursuer, let $\tau(g, v)$ be the time (possibly $+\infty$) when the game ends. Let

$$T(g) = \sup \{ \tau(g, v) \mid v \text{ is a strategy of the evader} \}$$

and let

$$T^* = \inf \{ T(g) \mid g \text{ is a strategy of the pursuer} \},$$

i.e.,

$$T^* = \inf_g \sup_v \tau(g, v).$$

We say that the game has a *solution* if there is a pursuit strategy g^* such that $T^* = T(g^*)$.

The main result of this paper (see §5) consists in showing that if $T^* < \infty$, then there exists a solution.

In §6 we consider the appropriateness of our model and discuss the relation of our results with the known [5], [6], [7] existence results on time optimal control. Section 2 deals with the postulates of the dynamics of the two players. In §3 and §4 we investigate the properties of the motion space and the strategy space, respectively.

2. Dynamics of the players. Instead of giving the dynamics of the two players by means of differential equations, we adopt the axioms of Roxin [1], [2]. The basic notion of these postulates is the attainability function $P(p_0, t_0, t)$ for the pursuer (or $E(e_0, t_0, t)$ for the evader), which represents the set attainable by the pursuer (or evader) at time t , while starting in state p_0 (or e_0) at time t_0 . There are two reasons for using this model. First of all, the mathematics is greatly simplified. More important is the belief that this simplified development enables us to distinguish the special problems arising in differential games, as opposed to optimal control.

The attainability functions have to satisfy the following axioms. (We only give the axioms for the pursuer since those for the evader can be obtained by replacing P by E and p by e .)

A1. $P(p_0, t_0, t_1)$ is defined for all p_0 in R^n and for all t_0 and t_1 with $0 \leq t_0 \leq t_1 < \infty$. For each value of the argument, $P(p_0, t_0, t)$ is a non-empty compact subset of R^n .

A2. For all p_0, t_0 , $P(p_0, t_0, t_0) = \{p_0\}$.

A3. For all p_0, t_0, t_1, t_2 with $t_0 \leq t_1 \leq t_2$,

$$P(p_0, t_0, t_2) = \bigcup_{p_1 \in P(p_0, t_0, t_1)} P(p_1, t_1, t_2).$$

A4. For fixed $p_0, t_0, P(p_0, t_0, t_1)$ is continuous in t_1 , i.e., for each p_0, t_0, t_1 and $\epsilon > 0$, there is a $\delta > 0$ such that¹

$$P(p_0, t_0, t_1) \subseteq P(p_0, t_0, t) + S_\epsilon$$

and

$$P(p_0, t_0, t') \subseteq P(p_0, t_0, t_1) + S_\epsilon$$

when

$$|t' - t_1| \leq \delta \quad \text{and} \quad t' \geq t_0.$$

A5. $P(p_0, t_0, t_1)$ is upper semicontinuous in the triple (p_0, t_0, t_1) , i.e., for each p_0, t_0, t_1 and $\epsilon > 0$ there is a $\delta > 0$ such that

$$P(p'_0, t'_0, t'_1) \subseteq P(p_0, t_0, t_1) + S_\epsilon$$

whenever²

$$|t'_1 - t_1| \leq \delta, \quad |t'_0 - t_0| \leq \delta, \quad |p'_0 - p_0| \leq \delta \quad \text{and} \quad t'_0 \leq t'_1.$$

3. Motions of the players. In this section we define the motions of the two players subject to the dynamical constraints of the previous section. The most important result (Theorem 3.1) is that the set of motions of each player over a fixed time interval $[0, T]$, $T < \infty$, is a compact set under a suitable topology. We assume throughout that each player starts from a fixed initial state.

DEFINITION 3.1.

a. A *motion* of the pursuer (or evader) is a mapping $u(\cdot)$ (or $v(\cdot)$) of a subinterval I of $[0, \infty)$ into R^n such that

(1) $0 \in I$ and $u(0) = p_0$ (or $v(0) = e_0$), where p_0 (or e_0) is the fixed initial state of the pursuer (or evader), and

(2) for t_0, t_1 in I with $t_0 \leq t_1$ we have

$$u(t_1) \in P(u(t_0), t_0, t_1) \quad (\text{or } v(t_1) \in E(v(t_0), t_0, t_1)).$$

We will say that the motion is *defined* on I .

b. Let $u(v)$ and $u_1(v_1)$ be motions defined on I and I_1 , respectively.

We say that $u_1(v_1)$ is a *prolongation* of $u(v)$ if

(1) $I \subseteq I_1$, and

(2) $u(t) = u_1(t)$ (or $v(t) = v_1(t)$) for t in I .

c. An *entire* motion is a motion defined on the entire interval $[0, \infty)$ of interest.

d. A pursuer (or evader) motion defined on $[0, T]$, $T < \infty$, will be

¹ Throughout S_ϵ represents the closed sphere in R^n of radius ϵ and center 0. Also, if A, B are subsets of R^n , then $A + B = \{a + b \mid a \in A, b \in B\}$.

² Throughout for $x \in R^n$, $|x|$ represents the Euclidean norm of x .

denoted by u_T (or v_T). The set of all such motions will be denoted by U_T (or V_T). An entire pursuer (or evader) motion will be denoted by \hat{u} (or \hat{v}), whereas the space of all such motions will be called \hat{U} (or \hat{V}).

Remark. In the main, we will be only interested in motions on a finite interval, i.e., in the spaces U_T and V_T .

For a proof of the following fact the reader is referred to Lemma 6.1 of [2].

LEMMA 3.1. *A motion defined on an interval I is necessarily a continuous mapping of I into R^n .*

DEFINITION 3.2. For $T < \infty$ let C_T denote the Banach space (see [3, pp. 261–281]) of all continuous mappings of the interval $[0, T]$ into R^n , where for $\xi \in C_T$ the norm of ξ is given by $\|\xi\| = \sup_{0 \leq t \leq T} |\xi(t)|$.

Because of Lemma 3.1, U_T and V_T can be considered to be subsets of C_T . We consider U_T and V_T as subspaces of C_T .

The next result follows directly from the axioms of §2.

LEMMA 3.2. *U_T and V_T are bounded subsets of C_T .*

LEMMA 3.3. *U_T and V_T are closed subsets of C_T .*

Proof. It is enough to prove the assertion for U_T since the proof for V_T is identical. Thus, let $\{u_{T,n}\}_{n=1}^\infty$ be a sequence in U_T which converges to an element ξ in C_T , i.e.,

$$(3.1) \quad \lim_n \sup_{0 \leq t \leq T} |u_{T,n}(t) - \xi(t)| = 0.$$

We have to show that $\xi \in U_T$. First of all, since $u_{T,n}(0) = p_0$ for each n , $\xi(0) = p_0$, so that by Definition 3.1a it remains to show that for all t_0, t_1 with $0 \leq t_0 \leq t_1 \leq T$,

$$\xi(t_1) \in P(\xi(t_0), t_0, t_1),$$

or, since $P(\xi(t_0), t_0, t_1)$ is closed, we have to show that for each $\epsilon > 0$,

$$(3.2) \quad \xi(t_1) \in P(\xi(t_0), t_0, t_1) + S_\epsilon.$$

Let $\epsilon > 0$. Because of A5 and (3.1), for sufficiently large n , say $n > n_1$,

$$u_{T,n}(t_1) \in P(u_{T,n}(t_0), t_0, t_1) \subseteq P(\xi(t_0), t_0, t_1) + S_{\epsilon/2}.$$

But again for large n , say $n > n_2$,

$$|\xi(t_1) - u_{T,n}(t_1)| \leq \frac{\epsilon}{2}.$$

Therefore, for $n > n_1 + n_2$, (3.2) is satisfied.

LEMMA 3.4. *U_T and V_T are equicontinuous subsets of C_T .*

Proof. Again we prove the assertion for U_T only. We have to show that for each $\epsilon > 0$ and for each $t \in [0, T]$, there is a $\delta > 0$, depending on ϵ ,

t such that

$$(3.2) \quad |u_T(t) - u_T(t')| \leq \epsilon$$

for all u_T in U_T and all t' in $[0, T]$ with $|t' - t| \leq \delta$. Suppose the assertion is false. Then there are $\epsilon > 0$, $t \in [0, T]$, and sequences $\{u_{T,n}\}_{n=1}^\infty \subseteq U_T$, $\{t_n\}_{n=1}^\infty \subseteq [0, T]$ such that

$$|t - t_n| \leq \frac{1}{n}, \quad |u_{T,n}(t_n) - u_{T,n}(t)| \geq \epsilon.$$

Taking subsequences, if necessary, we can assume that there are $x \in R^n$, $y \in R^n$ such that

$$(3.3) \quad u_{T,n}(t) \rightarrow x, \quad u_{T,n}(t_n) \rightarrow y \quad \text{as } n \rightarrow \infty,$$

so that

$$(3.4) \quad |x - y| \geq \epsilon.$$

Again, taking subsequences if necessary, we may assume that either (i) $t_n \leq t$ for all n or (ii) $t_n \geq t$ for all n .

Case (i). $t_n \leq t$ for all n .

By Axiom A5 of §2, using (3.3) we see that for large n ,

$$P(u_{T,n}(t_n), t_n, t) \subseteq P(y, t, t) + S_{\epsilon/2}.$$

But $t_n \leq t$ implies that $u_{T,n}(t) \in P(u_{T,n}(t_n), t_n, t)$ so that, for large n ,

$$|u_{T,n}(t) - y| \leq \frac{\epsilon}{2},$$

and hence $|x - y| \leq \epsilon/2$ which contradicts (3.4). Interchanging t_n and t in the above argument yields a contradiction of (3.4) for Case (ii) also. Hence, the assertion must be true.

The combination of Lemmas 3.2, 3.3 and 3.4 and the Ascoli-Arzelà theorem [3, p. 266] yields the following theorem.

THEOREM 3.1. U_T and V_T are compact subsets of C_T .

Remark. For an alternative proof of Theorem 3.1 see [2, Theorem 6.2].

4. Pursuit strategies and feasible pursuit strategies.

DEFINITION 4.1.

a. A *pursuit strategy* is a mapping $g_T : V_T \rightarrow U_T$ such that, if v_T and v_T' are in V_T and

$$v_T(\tau) = v_T'(\tau) \quad \text{for } 0 \leq \tau \leq t,$$

then

$$g_T v_T(\tau) = g_T v_T'(\tau) \quad \text{for } 0 \leq \tau \leq t.$$

We say that g_T is *defined* on $[0, T]$.

b. Let G_T denote the set of all strategies defined on $[0, T]$.

DEFINITION 4.2.

a. Let $F(V_T, U_T)$ be the space of all mappings η from V_T into U_T . We give $F(V_T, U_T)$ the topology of pointwise convergence (see [4, p. 217]). Thus a net $\{\eta_\alpha\} \subseteq F(V_T, U_T)$ converges to an element η of $F(V_T, U_T)$ if and only if $\eta_\alpha(v_T)$ converges to $\eta(v_T)$ in U_T for each v_T in V_T .

b. We can consider G_T as a subset of $F(V_T, U_T)$ and give G_T the relative topology.

DEFINITION 4.3. Let Θ be a fixed closed subset of $[0, \infty) \times R^n$. Θ is called the *endzone*.

DEFINITION 4.4.

a. A pursuit strategy $g_T \in G_T$ is said to be *feasible* if for each v_T in V_T

$$(\tau, g_T v_T(\tau) - v_T(\tau)) \in \Theta$$

for some $\tau, 0 \leq \tau \leq T$.

b. We will denote a feasible pursuit strategy defined on $[0, T]$ by f_T , and the set of all feasible strategies defined on a fixed interval $[0, T]$ by F_T . We consider F_T as a subspace of $F(V_T, U_T)$.

Thus a feasible pursuit strategy f_T is a strategy which guarantees the pursuer that the game will end in time at most T , independent of the strategy used by the evader.

LEMMA 4.1. F_T is a closed subset of $F(V_T, U_T)$.

Proof. Let $\{f_{T,\alpha}\}$ be a net in F_T which converges to an η of $F(V_T, U_T)$. We must show that $\eta \in F_T$.

By the definition of convergence in $F(V_T, U_T)$ we have that for each v_T in V_T

$$(4.1) \quad f_{T,\alpha}(v_T) \rightarrow \eta(v_T) \quad \text{in } U_T.$$

First of all, let v_T and v_T' be in V_T such that $v_T(\tau) = v_T'(\tau)$ for $0 \leq \tau \leq t$. Since $f_{T,\alpha}$ is a pursuit strategy for each α , we must have $f_{T,\alpha} v_T(\tau) = f_{T,\alpha} v_T'(\tau)$ for $0 \leq \tau \leq t$ so that (4.1) implies that $\eta v_T(\tau) = \eta v_T'(\tau)$ for $0 \leq \tau \leq t$. Therefore η is certainly a pursuit strategy.

Now let v_T be a fixed element of V_T . For each α there is a $\tau_\alpha, 0 \leq \tau_\alpha \leq T$, such that (see Definition 4.4a)

$$(\tau_\alpha, f_{T,\alpha} v_T(\tau_\alpha) - v_T(\tau_\alpha)) \in \Theta.$$

Taking subnets if necessary, we can assume that τ_α converges to τ^* for some τ^* with $0 \leq \tau^* \leq T$. We next show that

$$(4.2) \quad f_{T,\alpha} v_T(\tau_\alpha) - v_T(\tau_\alpha) \rightarrow \eta v_T(\tau^*) - v_T(\tau^*),$$

$$\begin{aligned}
 & |\eta v_T(\tau^*) - v_T(\tau^*) - f_{T,\alpha} v_T(\tau_\alpha) - v_T(\tau_\alpha)| \\
 (4.3) \quad & \leq |\eta v_T(\tau^*) - \eta v_T(\tau_\alpha)| + |\eta v_T(\tau_\alpha) - f_{T,\alpha} v_T(\tau_\alpha)| \\
 & \quad + |v_T(\tau^*) - v_T(\tau_\alpha)|.
 \end{aligned}$$

Now, v_T , ηv_T , $f_{T,\alpha} v_T$ are continuous functions on $[0, T]$; furthermore, convergence in U_T means uniform convergence over $[0, T]$ so that from (4.1) we see that each of the three terms in (4.3) converges to zero. Hence (4.2) is true. Therefore,

$$(4.4) \quad (\tau_\alpha, f_{T,\alpha} v_T(\tau_\alpha) - v_T(\tau_\alpha)) \rightarrow (\tau^*, \eta v_T(\tau^*) - v_T(\tau^*))$$

in $[0, \infty) \times R^n$.

But each term in the left-hand side of (4.4) belongs to Θ and Θ is a closed subset so that the right-hand term also belongs to Θ . This proves that η is a feasible pursuit strategy (see Definition 4.4a), i.e., $\eta \in F_T$, so that F_T is closed.

THEOREM 4.1. F_T is a compact subset of $F(V_T, U_T)$.

Proof. The topology on $F(V_T, U_T)$ is the product topology on the product space of $|V_T|$ copies³ of the space U_T . But by Theorem 3.1, U_T is compact. By the Tychonoff product theorem [4, p. 143], $F(V_T, U_T)$ is compact. By Lemma 4.1, F_T is a closed subset of $F_T(V_T, U_T)$ so that F_T is compact.

5. Existence of solutions. Suppose that there is a T , $0 < T < \infty$, such that F_T is nonempty, i.e., suppose that there exists a feasible pursuit strategy.

DEFINITION 5.1. For each v_T in V_T and each f_T in F_T let $\tau(f_T, v_T)$ be the smallest number τ such that $(\tau, f_T v_T(\tau) - v_T(\tau)) \in \Theta$.

Remark. Since Θ is a closed subset of $[0, \infty) \times R^n$ and $f_T v_T(\tau) - v_T(\tau)$ is a continuous function on $[0, T]$, $\tau(f_T, v_T)$ is well-defined.

DEFINITION 5.2.

a. For each f_T in F_T let

$$T(f_T) = \sup_{v_T \in V_T} \tau(f_T, v_T).$$

b. Let

$$T^* = \inf_{f_T \in F_T} T(f_T).$$

c. We say that the game has a *well solution* if there is an f_T^* in F_T such that

³ $|V_T|$ denotes the cardinality of the set V_T .

$$T(f_T^*) = T^*.$$

We will call f_T^* a well solution of the game.

LEMMA 5.1. *For fixed v_T , $\tau(f_T, v_T)$ is a lower semicontinuous function [4, pp. 101–102] on F_T , i.e., for each a the set $\{f_T \mid \tau(f_T, v_T) \leq a\}$ is a closed subset of F_T .*

Proof. Let a be fixed. Let $\{f_{T,\alpha}\} \subseteq F_T$ be a net such that

(i) $\tau(f_{T,\alpha}, v_T) \leq a$ for each α , and

(ii) $f_{T,\alpha}$ converges to an element f_T of F_T .

We have to show that $\tau(f_T, v_T) \leq a$, i.e., we must show that there is a τ , $0 \leq \tau \leq a$, such that

$$(5.1) \quad (\tau, f_T v_T(\tau) - v_T(\tau)) \in \Theta.$$

Because of (i), for each α there is a τ_α with $0 \leq \tau_\alpha \leq a$ such that

$$(\tau_\alpha, f_{T,\alpha} v_T(\tau_\alpha) - v_T(\tau_\alpha)) \in \Theta.$$

Taking subnets, if necessary, we can assume that $\tau_\alpha \rightarrow \tau^*$ and $0 \leq \tau^* \leq a$. But then the same argument as in Lemma 4.1 shows that (5.1) is satisfied with $\tau = \tau^*$.

LEMMA 5.2. *$T(f_T)$ is a lower semicontinuous function on F_T .*

Proof. Let a be a fixed real number and let $\{f_{T,\alpha}\} \subseteq F_T$ be a set such that

(i) $T(f_{T,\alpha}) \leq a$ for each α , and

(ii) $f_{T,\alpha}$ converges to an element f_T of F_T .

We must show that $T(f_T) \leq a$. Let v_T be an arbitrary element of V_T . Because of (i) above,

$$\tau(f_{T,\alpha}, v_T) \leq a \quad \text{for each } \alpha.$$

But then by Lemma 5.1, $\tau(f_T, v_T) \leq a$, so that

$$T(f_T) = \sup_{v_T \in V_T} \tau(f_T, v_T) \leq a.$$

THEOREM 5.1. *If there exists a feasible pursuit strategy, then there exists a well solution to the game.*

Proof. By the hypothesis there is a T , $0 \leq T < \infty$, such that F_T is nonempty. By Theorem 4.1, F_T is a nonempty compact set. By Lemma 5.2, $T(f_T)$ is a lower semicontinuous function on F_T and hence it has a minimum at some point f_T^* of F_T . Clearly, f_T^* is a well solution to the game.

6. Discussion of the model and relation with time-optimal control problems. Suppose that the dynamics of the players are given by differential equations instead of via the axioms of §2. For example, let the pursuer dynamics be given by

$$(S) \quad \dot{p}(t) = f(p(t), \sigma(t), t), \quad p(0) = p_0,$$

where $\sigma(t) \in \Sigma \subseteq R^m$ is the control vector. Suppose that Σ is bounded, and f satisfies enough conditions to insure uniqueness and boundedness of the solution for each measurable control function $\sigma(t)$ with range in Σ . We can define the attainability function $P(p_0, t_0, t)$ for this differential system (S) to be the set of all states p which can be reached at time t , starting in p_0 at time t_0 , by using an admissible control. Then under mild conditions on f , $P(p_0, t_0, t)$ is bounded for each value of its argument and the attainability function satisfies Axioms A2–A5. However, in general $P(p_0, t_0, t)$ is not a closed subset of R^n (see [8]). Let us assume that conditions are imposed on f (see [5]–[8]) such that $P(p_0, t_0, t)$ is closed. Now the system (S) has a well-defined notion of *trajectory* which is any solution of (S) arising from a measurable control. The attainability function P , on the other hand, gives rise to the concept of *motion* as in §3. A little reflection shows that every trajectory is a motion. However, the converse is not true in general. The precise relations between the set of trajectories of (S) and the set of “derived motions” will be investigated in another paper. This relationship is very similar to the one between the “original curves” and the “relaxed curves” discussed by J. Warga [9]. In [5]–[8], varying sets of sufficient conditions are imposed on f (see (S)) which insure that (a) the attainable set is closed, and (b) every motion is a trajectory. If any of these sets of conditions are satisfied by (S), Theorem 5.1 tells us that if there exists a feasible solution to the time optimal problem, there exists an optimal solution.

7. Summary and conclusion. In this paper we have considered the problem of the existence of a “min-sup strategy” to a pursuit-evasion game. The dynamics of the players have been modeled by a general dynamical system (as developed by Zubov, Roxin, and others) instead of a differential system, the purpose being to clarify the problems arising in the competitive situation of a game as distinct from an optimal control problem. Usually in game theory, interest centers around the existence of a more symmetric solution, i.e., on both “minimax” and “max-min” strategies. The existence of such solutions seems to the author to be extremely unlikely in a differential game, except under very restrictive conditions. In any case, it is hoped that the results of this paper will evoke interest among both game theoreticians and people working in the theory of optimal control.

REFERENCES

- [1] E. ROXIN, *Dynamical systems with inputs*, Proceedings of the Symposium on System Theory, Polytechnic Institute of Brooklyn, New York, 1965, pp. 105–114.
- [2] ———, *Stability in general control systems*, J. Differential Equations, 1 (1965), pp. 115–150.

- [3] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part I: General Theory*, Interscience, New York, 1964.
- [4] J. L. KELLEY, *General Topology*, D. Van Nostrand, Princeton, 1957.
- [5] E. ROXIN, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109–119.
- [6] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
- [7] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.
- [8] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.
- [9] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111–128.

OPTIMAL CONTROL IN A LINEAR SYSTEM UNDER CONFLICT*

V. B. GINDES†

Abstract. The article considers the problem of optimal control in a linear system by two players who have conflicting goals. The state of the system is optimized at a given instant of time under constraints on the state coordinates at discrete instants of time. The present problem was formulated by R. Gabasov and F. M. Kirillova.

1. Statement of the problem. The plant to be controlled is described by the linear differential equation

$$(1) \quad \dot{x} = Ax + Bu + Cv$$

with the initial conditions $x(t_0) = x_0$, where $x = x(t)$ is the n -dimensional state-coordinate vector of the plant; $u = u(t)$ and $v = v(t)$ are the r - and p -dimensional control vectors of the first and second player, respectively; the coefficient matrices A , B , C are continuous functions of time. Clearly the state of the plant at the instant t is completely determined by each player's choice of the control functions $u(\tau)$ and $v(\tau)$, $t_0 \leq \tau \leq t$, from corresponding sets U and V of admissible controls. Below we shall suppose that U and V are classes of piecewise-continuous vector-valued functions satisfying, respectively, the conditions

$$(2) \quad \|u\| \leq 1 \quad \text{and} \quad \|v\| \leq 1,$$

where the symbol $\|w\|$ denotes the norm of the element w in the space L_p .

The players choose the controls u and v over the whole control process interval $t_0 \leq t \leq T$ which is known a priori; moreover, Player II chooses his control first, while Player I makes his choice knowing what his opponent has chosen. This situation corresponds to the programmed control of a plant in the interval $[t_0, T]$ in which Player I when choosing his own program, the function $u(t)$, $t_0 \leq t \leq T$, knows the program, $v(t)$, $t_0 \leq t \leq T$, of Player II.

Furthermore, the players are at cross purposes: Player I strives to bring the plant at a specified instant θ , $t_0 < \theta \leq T$, nearer (in the sense of the norm in the state space X_n of system (1)) to a given point c_θ of the state space, while Player II, on the contrary, strives to increase this distance as much as possible. Moreover, in the process of being controlled the normal operating conditions of the plant should be satisfied, by which we mean

* Originally published in *Izv. Vyssh. Uchebn. Zaved. Matematika*, No. 3, 52 (1966), pp. 39–44. Submitted on January 19, 1965, for publication.

This translation into English has been prepared by N. H. Choksy, and was supported in part by a grant-in-aid from the National Science Foundation.

† S. M. Kirov Ural Polytechnic Institute, Sverdlovsk 2, USSR.

the following: at assigned control instants t_k , $k = 1, 2, \dots, N$, $t_0 < t_1 < t_2 < \dots < t_N \leq T$, the state of the plant in the state space should not leave specified neighborhoods of fixed points c_k , $k = 1, 2, \dots, N$.

We introduce the notations:

$$(3) \quad z(t_k) = x(t_k) - c_k, \quad z(\theta) = x(\theta) - c_\theta, \quad \rho = \|z(\theta)\|.$$

The quantity to be optimized is the distance of the plant state vector from the point c_θ ; this distance is a functional ρ whose value is determined by the control functions $\rho = \rho(u, v)$. By what was stated before, the problem consists of finding the control functions u^0 and v^0 (the optimal controls) of both players and the quantity ρ^0 (the optimal distance) such that

$$(4) \quad \rho^0 = \rho(u^0, v^0) = \min_{u \in U} \rho(u, v^0) = \max_{v \in V} \min_{u \in U} \rho(u, v)$$

under the conditions

$$(5) \quad \|z(t_k)\| = \|x(t_k) - c_k\| \leq \epsilon_k, \quad \epsilon_k > 0, \quad k = 1, 2, \dots, N.$$

The stated problem can be related to a pursuit problem. References [1]–[5] have investigated pursuit problems in which the pursuit time is optimized. A pursuit problem with a fixed time has been considered in [6], [7]. The methods proposed in [6] are utilized to solve the formulated problem with bounded state coordinates.

2. Solution of the problem. By the Cauchy formula the solution of (1) has the form

$$(6) \quad x(t) = F(t, t_0)x_0 + \int_{t_0}^t F(t, \tau)B(\tau)u(\tau) d\tau + \int_{t_0}^t F(t, \tau)C(\tau)v(\tau) d\tau,$$

where $dF(t, t_0)/dt = AF(t, t_0)$, $F(t_0, t_0) = I$, the identity. We introduce the notations

$$F(t_k, t_0)x_0 = b_k,$$

$$\int_{t_0}^{t_k} F(t_k, \tau)B(\tau)u(\tau) d\tau = S_k u, \quad \int_{t_0}^{t_k} F(t_k, \tau)C(\tau)v(\tau) d\tau = -P_k v;$$

the analogous quantities at the instant θ are denoted by b_θ , $S_\theta u$, $-P_\theta v$.

The elements b_i , $S_i u$, $P_i v$, $i = 1, \dots, N$, θ , belong to the space X_n , while S_i and P_i can be interpreted as linear operators mapping the elements u and v of functional spaces into elements of the finite-dimensional state space X_n . The functions $u = u(t)$ and $v = v(t)$ are defined on the interval $[t_0, T]$, while the elements of the dual spaces are defined on the

same time interval by the formulas

$$\begin{aligned} S_k(t) &= \begin{cases} F(t_k, t)B(t), & t_0 \leq t \leq t_k, \\ 0, & t_k < t \leq T; \end{cases} \\ P_k(t) &= \begin{cases} -F(t_k, t)C(t), & t_0 \leq t \leq t_k, \\ 0, & t_k < t \leq T; \end{cases} \\ S_\theta(t) &= \begin{cases} F(\theta, t)B(t), & t_0 \leq t \leq \theta, \\ 0, & \theta < t \leq T; \end{cases} \\ P_\theta(t) &= \begin{cases} -F(\theta, t)C(t), & t_0 \leq t \leq \theta, \\ 0, & \theta < t \leq T. \end{cases} \end{aligned}$$

In (6) let $t = t_k$, $k = 1, 2, \dots, N$, and also let $t = \theta$. In terms of the notation we have introduced we have

$$\begin{aligned} x(t_k) &= b_k + S_k u - P_k v, \quad k = 1, 2, \dots, N, \\ x(\theta) &= b_\theta + S_\theta u - P_\theta v. \end{aligned}$$

Substituting these expressions into (3) we arrive at a system of operator equations

$$(7) \quad \begin{aligned} S_k u - P_k v - z(t_k) &= c_k - b_k, \quad k = 1, 2, \dots, N, \\ S_\theta u - P_\theta v - z(\theta) &= c_\theta - b_\theta. \end{aligned}$$

We next introduce the vector space W of dimension $m = n(N + 1)$. In the space W , (7) can be represented by the single operator equation

$$(8) \quad Su - Pv - \sum_{k=1}^N R_k z(t_k) - R_\theta z(\theta) = d,$$

where

$$\begin{aligned} S &= \begin{pmatrix} S_1 \\ \vdots \\ S_N \\ S_\theta \end{pmatrix}, \quad P = \begin{pmatrix} P_1 \\ \vdots \\ P_N \\ P_\theta \end{pmatrix}, \quad R_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ E \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \left. \vphantom{\begin{bmatrix} 0 \\ \vdots \\ 0 \\ E \\ 0 \\ \vdots \\ 0 \end{bmatrix}} \right\} (k-1) \text{ null matrices} \\ R_\theta &= \begin{pmatrix} 0 \\ \vdots \\ 0 \\ E \end{pmatrix}, \quad d = \begin{pmatrix} c_1 - b_1 \\ \vdots \\ c_N - b_N \\ c_\theta - b_\theta \end{pmatrix}. \end{aligned}$$

Thus, the problem stated in §1 is reduced to that of finding, from among the solutions of (8) which satisfy the constraint relations (2) and (5), that solution $\{u^0, v^0, z^0(t_k), z^0(\theta)\}$ for which the quantity $\rho^0 = \|z^0(\theta)\|$ satisfies condition (4).

Let us assume that Player II has chosen the control $v \in V$. We consider the problem of finding the quantity $\min_{u \in U} \rho(u, v)$. Let the number $\epsilon_\theta \geq 0$ be given.

FUNDAMENTAL ASSERTION. *For fixed v and under the conditions*

$$(9) \quad \|u\| \leq 1, \quad \|z(t_k)\| \leq \epsilon_k, \quad \|z(\theta)\| \leq \epsilon_\theta,$$

(8) has a solution if and only if the inequality

$$(10) \quad \|S^*g\| - g(Pv) + \sum_{k=1}^N \|R_k^*g\| \epsilon_k + \|R_\theta^*g\| \epsilon_\theta \geq g(d)$$

is satisfied for every vector $g \in W^*$.

Proof. Necessity. If $\bar{u}, \bar{z}(t_k), \bar{z}(\theta), k = 1, 2, \dots, N$, satisfy (8) and conditions (9), then, for any $g \in W^*$,

$$g \left[Su - Pv - \sum_{k=1}^N R_k \bar{z}(t_k) - R_\theta \bar{z}(\theta) \right] = g(d),$$

$$(S^*g, \bar{u}) - g(Pv) - \sum_{k=1}^N (R_k^*g, \bar{z}(t_k)) - (R_\theta^*g, \bar{z}(\theta)) = g(d),$$

where the asterisk denotes formation of the adjoint operator. Since

$$(S^*g, \bar{u}) \leq \|S^*g\| \cdot \|\bar{u}\| \leq \|S^*g\|,$$

$$-(R_k^*g, \bar{z}(t_k)) \leq \|R_k^*g\| \cdot \|\bar{z}(t_k)\| \leq \|R_k^*g\| \epsilon_k, \quad k = 1, 2, \dots, N,$$

$$-(R_\theta^*g, \bar{z}(\theta)) \leq \|R_\theta^*g\| \cdot \|\bar{z}(\theta)\| \leq \|R_\theta^*g\| \epsilon_\theta,$$

condition (10) is satisfied. This proves its necessity.

Sufficiency. Let inequality (10) be satisfied for any $g \in W^*$. Consider the closed convex set $L \subset W$ of elements l of the form

$$l = Su - Pv - \sum_{k=1}^N R_k z(t_k) - R_\theta z(\theta),$$

where $u, z(t_1), z(t_2), \dots, z(t_N), z(\theta)$ take on all possible values satisfying conditions (9). Let us assume that the problem (8)–(9) has no solution. This means that the point d does not belong to the set L . In this case, by the separation theorem for closed convex sets there exist a linear functional $\bar{g} \in W^*$ and a number α such that $\bar{g}(l) < \alpha$ for all $l \in L$ and $\bar{g}(d) > \alpha$,

i.e., $\max_{l \in L} \bar{g}(l) < \bar{g}(d)$. Hence,

$$\max_{u, z(t_k), z(\theta)} \bar{g} \left[Su - Pv - \sum_{k=1}^N R_k z(t_k) - R_\theta z(\theta) \right] < \bar{g}(d),$$

$$\|S^* \bar{g}\| - g(Pv) + \sum_{k=1}^N \|R_k^* \bar{g}\| \epsilon_k + \|R_\theta^* \bar{g}\| \epsilon_\theta < \bar{g}(d), \quad \bar{g} \in W^*,$$

which contradicts inequality (10). The assertion is proved.

Let $\|R_\theta^* g\| \neq 0$. Then from (10) it follows that

$$\epsilon_0 \geq \frac{g(d) - \|S^* g\| + g(Pv) - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}$$

for all $g \in W^*$, or

$$\epsilon_\theta \geq \max_{g \in W^*} \frac{g(d) - \|S^* g\| + g(Pv) - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}.$$

Hence,

$$(11) \quad \min_{u \in U} \rho(u, v) = \min_{g \in W^*} \frac{g(d) - \|S^* g\| + g(Pv) - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}.$$

From (10) we see that when $\|R_\theta^* g\| = 0$, the numerator in (11) is nonpositive and, therefore, the maximum in (11) is reached when $\|R_\theta^* g\| \neq 0$. From (11) we have

$$\rho^0 = \max_{v \in V} \min_{u \in U} \rho(u, v)$$

$$= \max_{v \in V} \max_{g \in W^*} \frac{g(d) - \|S^* g\| + g(Pv) - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}.$$

Using permutability of the "maximum" operation we obtain

$$\rho^0 = \max_{g \in W^*} \max_{v \in V} \frac{g(d) - \|S^* g\| + g(Pv) - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}.$$

Since $\max_{v \in V} g(Pv) = \|P^* g\|$, we have finally

$$(12) \quad \rho^0 = \max_{g \in W^*} \frac{g(d) - \|S^* g\| + \|P^* g\| - \sum_{k=1}^N \|R_k^* g\| \epsilon_k}{\|R_\theta^* g\|}.$$

Let $g^0 \in W^*$ be some vector for which the maximum in (12) is reached, i.e.,

$$\rho^0 = \frac{g^0(d) - \|S^*g^0\| + \|P^*g^0\| - \sum_{k=1}^N \|R_k^*g^0\| \epsilon_k}{\|R_\theta^*g^0\|}.$$

Then, the optimal controls will satisfy the maximum principle

$$(13) \quad \begin{aligned} (S^*g^0, u^0) &= \max_{u \in U} (S^*g^0, u) = \|S^*g^0\|, \\ (P^*g^0, v^0) &= \max_{v \in V} (P^*g^0, v) = \|P^*g^0\|. \end{aligned}$$

Thus, the solution of the optimization problem stated in §1 has been reduced to the finite-dimensional problem (12). For smooth norms the problem (12) can be solved by the usual methods of analysis. If

$$\begin{aligned} \|u\| &= \max_i \max_{t_0 \leq t \leq T} |u_i(t)|, & i &= 1, 2, \dots, r, \\ \|v\| &= \max_i \max_{t_0 \leq t \leq T} |v_i(t)|, & i &= 1, 2, \dots, p, \end{aligned}$$

then from (13) follows

$$\begin{aligned} u_i^0(t) &= \text{sign} [S^*g^0]_i(t), & i &= 1, 2, \dots, r, \\ v_i^0(t) &= \text{sign} [P^*g^0]_i(t), & i &= 1, 2, \dots, p. \end{aligned}$$

Remark 1. The relations (13) establish the connection between the method described here and Pontryagin's maximum principle. The vector g^0 corresponds to the vector ψ_0 , the initial conditions for the adjoint system [1]. Expression (12) is a supplementary relation involving this vector.

Remark 2. The method described can also be applied, without change, in the case when not all the components of the state vector are bounded at the control instants, and also when optimization is effected in some linear subspace of the state space X_n . For example, for a mechanical system, only the spatial position of the plant need be optimized, while the components of the vector x corresponding to the velocities may not be taken into account. Conversely, possibly only the velocities need be bounded at the control instants. Similar special cases may arise because of incomplete observability of the plant at various instants of time. Here the vectors $z_1, z_2, \dots, z_N, z_\theta$ will have the dimensions of the corresponding subspaces of the state space, and the dimension of space W will be changed; but from the mathematical point of view no complications arise here since when proving the fundamental assertion we used only the linearity of the operators S, P, R_k, R_θ , while the nature of the spaces on whose elements they act is immaterial.

Remark 3. In order for the problem stated in §1 to have a solution it is necessary to satisfy the condition

$$\|S^*g\| + \|P^*g\| + \sum_{k=1}^N \|R_k^*g\| \epsilon_k \geq g(d)$$

for all $g \in W^*$ for which $\|R_\theta^*g\| = 0$. Indeed, if $\|R_\theta^*g'\| = 0$ for some $g' \in W^*$ and if $\|S^*g'\| + \|P^*g'\| + \sum_{k=1}^N \|R_k^*g'\| \epsilon_k < g'(d)$, then, obviously, for g' condition (10) is not satisfied for any choice of ϵ_θ and $v \in V$. This is an indication of the inconsistency of conditions (2)–(5), i.e., of the fact that it is impossible to drive the system through the control neighborhoods by means of admissible controls. In this case the optimization problem is meaningless.

Remark 4. The problem considered can be interpreted as an approximation to the problem of optimal control with bounded state coordinates. The difference consists in the finiteness of the number of control instants. When the control points are sufficiently “densely” distributed, then, by taking the continuity of the state trajectories into account, we can speak of approximating the continuous bounds by bounds at isolated time instants. However, it should be noted that an increase in the number of control points implies an increase in the dimension of space W , which might be a practical restriction on the applicability of the method described to the solution of actual optimal control problems.

The author is indebted to R. Gabasov, F. M. Kirillova and Yu. I. Alimov for their attention to the present work.

REFERENCES

- [1] I. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [2] D. L. KELENDZHERIDZE, *Theory of an optimal pursuit strategy*, Soviet Math. Dokl., 2 (1961), pp. 654–656.
- [3] N. N. KRASOVSKII, *On a problem in tracking*, J. Appl. Math. Mech., 26 (1962), pp. 314–335.
- [4] ———, *On a problem of tracking*, Ibid., 27 (1963), pp. 363–377.
- [5] L. S. GNOENSKII, *On the tracking problem*, Ibid., 26 (1962), pp. 1451–1460.
- [6] R. GABASOV AND F. M. KIRILLOVA, *The solution of some problems in the theory of optimal processes*, Automat. Remote Control, 25 (1964), pp. 945–955.
- [7] V. B. GINDES, *Optimal processes for two controlled systems*, Proceedings of the Third Siberian Conference on Mathematics and Mechanics, Tomsk, 1964, pp. 260–261.

A SYSTEM THEORY CRITERION FOR POSITIVE REAL MATRICES*

B. D. O. ANDERSON†

1. Introduction. The concept of a positive real function is now an old one of network theory, and more recently the concept has been usefully employed in other system theoretic investigations, such as in the development of the Popov criterion for the stability of a feedback system containing a single memoryless nonlinearity [1]. In view of this and other connections between network and control theory, it seems possible that the concept of a *positive real matrix* could be employed fruitfully in control systems investigations; to assist in such investigations this paper discusses positive real matrices from a system theory viewpoint.

An $n \times n$ matrix $Z(\cdot)$ of functions of a complex variable is called positive real if [2]

- (i) $Z(s)$ has elements which are analytic for $\text{Re } s > 0$;
- (ii) $Z^*(s) = Z(s^*)$ for $\text{Re } s > 0$;
- (iii) $Z'(s^*) + Z(s)$ is nonnegative definite for $\text{Re } s > 0$.

Here the superscript asterisk denotes complex conjugation; the prime denotes matrix transposition.

This paper is concerned with developing a criterion for a matrix of *rational* functions to be positive real. The criterion is a systems theoretic one, in the sense that it is formulated in terms of the parameters of a control system realization of the matrix. Such a result has been developed elsewhere for the scalar case and applied to the development of the Popov criterion [1]. A generalization of the criterion for a scalar function to be positive real has been stated for a subclass of positive real matrices, namely, those which are zero at $s = \infty$, [3]. In this reference, no proof was given; an outline proof was suggested, but the filling in of the details presented more difficulties than suspected, and a correct proof has not been available hitherto. Similar ideas are discussed in a paper of Popov [4].

After a review of some concepts associated with system realizations and some mathematical preliminaries in §2, we present in §3 the statement and proof of the positive real criterion. Initially, we consider matrices whose elements have no poles on the imaginary axis and which are zero at $s = \infty$;

* Received by the editors March 7, 1966, and in revised form November 15, 1966.

† Formerly with Stanford Electronics Laboratories, Stanford University, Stanford, California. Now with the Department of Electrical Engineering, University of Newcastle, New South Wales, Australia. This work was supported by the National Science Foundation under Grant GK 237. The author wishes also to acknowledge the support of the Services Canteens Trust Fund, an agency of the Australian Government, and the United States Education Foundation in Australia for a Fulbright grant.

then matrices with poles only on the imaginary axis; and, finally, general positive real matrices (i.e., those which are not necessarily zero at $s = \infty$, and which have imaginary axis poles permitted). This result appears as Theorem 3, and includes all preceding results as special cases.

Of course Theorem 3 alone could be proved, but the motivation for its proof would then be obscure. As it stands, it is a natural outgrowth of the earlier, more motivated, theorems.

Notation in the paper is straightforward: capital letters will be used for matrices, small letters for vectors. Other symbols will be explained as required.

2. Mathematical and system theoretic preliminaries. In this section we review some of the concepts associated with *linear time-invariant dynamical systems*; these concepts will appear in the discussion of positive real matrices. We shall also state some mathematical results to be used in the sequel.

We assume that $M(s)$ is an $m \times n$ matrix of rational functions, with $M(\infty) = 0$. Then a triple $\{F, G, H\}$ is termed a *realization* of M (see [5], [6]) if

$$(1) \quad M(s) = H'(sI - F)^{-1}G.$$

This is because $M(s)$ is the transfer function matrix relating an input vector u to an output vector y in the following state-space representation of M :

$$(2a) \quad \dot{x} = Fx + Gu,$$

$$(2b) \quad y = H'x.$$

Here x is a p -vector, the state; u is an n -vector, the input; y is an m -vector, the output; F is $p \times p$, G is $p \times n$, and H is $p \times m$.

For a given $M(s)$, there exist infinitely many sets $\{F, G, H\}$ constituting a realization [5]. Clearly there must be a *minimal dimension* which F may have; any realization incorporating F of minimal dimension is termed a *minimal realization*.

LEMMA 1 [7]. Let $\{F_1, G_1, H_1\}$ and $\{F_2, G_2, H_2\}$ be two minimal realizations of $M(s)$. Then there exists a nonsingular T such that

$$(3a) \quad F_2 = TF_1T^{-1},$$

$$(3b) \quad G_2 = TG_1,$$

$$(3c) \quad H_2 = (T')^{-1}H_1.$$

Conversely, if $\{F_1, G_1, H_1\}$ is minimal and T is nonsingular, $\{F_2, G_2, H_2\}$ as given by (3a), (3b) and (3c) is minimal.

The *dimension of a minimal realization* is the dimension of a minimal F matrix and is termed the *degree* of $M(s)$; it is, naturally, a positive integer number uniquely determined by $M(s)$, written $\delta[M(s)]$ or $\delta[M]$. Actually, numerous definitions of degree have appeared over the last fifteen years [8], [9], [10], but recently these have been reconciled with one another [11].

We shall have occasion to use several properties as set out below.

LEMMA 2 [8], [11]. *If the elements of $M_1(\cdot)$ and $M_2(\cdot)$ have no poles in common,*

$$(4) \quad \delta[M_1 + M_2] = \delta[M_2].$$

LEMMA 3 [8], [11].

$$(5) \quad \delta[M_1 M_2] \leq \delta[M_1] + \delta[M_2].$$

LEMMA 4. *If $M_1(\cdot)$ is $n \times r$, $M_2(\cdot)$ is $r \times n$, $r \leq n$, the elements of M_1 and M_2 have no common poles, $\text{rank } M_1(s_0) = r$ at any pole s_0 of an element of M_2 , and $\text{rank } M_2(\hat{s}_0) = r$ at any pole \hat{s}_0 of an element of M_1 , then*

$$(6) \quad \delta[M_1 M_2] = \delta[M_1] + \delta[M_2].$$

Proof. We use the Smith-McMillan decomposition [8], [11] for M_1 and M_2 . We have $M_1 = A_1 \Gamma_1 B_1$ and $M_2 = A_2 \Gamma_2 B_2$, where A_1, B_2 are $n \times n$ polynomial matrices with constant determinant; B_1, A_2 are $r \times r$ polynomial matrices with constant determinant; Γ_1 has its first r rows given by $\text{diag} [\epsilon_1/\psi_1, \dots, \epsilon_r/\psi_r]$ and its last $n - r$ rows zero; Γ_2 has its first r columns given by $\text{diag} [\lambda_1/\mu_1, \dots, \lambda_r/\mu_r]$ and its last $n - r$ columns zero. The ϵ_i , etc., are polynomials. Denote by $\hat{\Gamma}_1$ the first r rows of Γ_1 and by $\hat{\Gamma}_2$ the first r columns of Γ_2 . Then at poles of the elements of $\hat{\Gamma}_1$, $\hat{\Gamma}_2$ is a non-singular matrix, and vice versa.

This implies (see [8, §§5.26 and 5.4]) that $\delta[\hat{\Gamma}_1 B_1 A_2 \hat{\Gamma}_2] = \delta[\hat{\Gamma}_1] + \delta[\hat{\Gamma}_2]$. The matrix $\Gamma_1 B_1 A_2 \Gamma_2$ is simply $\hat{\Gamma}_1 B_1 A_2 \hat{\Gamma}_2$ with rows and columns of zeros added, and thus [8, §5.45], $\delta[\Gamma_1 B_1 A_2 \Gamma_2] = \delta[\hat{\Gamma}_1 B_1 A_2 \hat{\Gamma}_2]$. Finally, by [8, §5.16], we see that $\delta[M_1 M_2] = \delta[\Gamma_1 B_1 A_2 \Gamma_2]$, and $\delta[M_1] = \delta[\hat{\Gamma}_1]$, $\delta[M_2] = \delta[\hat{\Gamma}_2]$, from which the result immediately follows.

The next lemma, though purely algebraic in character, is of great use in studying the stability of linear systems, being originally due to Lyapunov. For a discussion and proof of the first part of the result, see [12], and for the second part, see [1].

LEMMA 5. *Let F be a $p \times p$ matrix with eigenvalues all possessing negative real parts. Then to each $p \times q$ matrix L (q arbitrary) there corresponds a unique symmetric nonnegative definite solution P to the equation*

$$(7) \quad PF + F'P = -LL'.$$

Moreover if $[F, L']$ is completely observable [5], or equivalently, if

$L' \exp (Ft)x = 0$ for all t implies $x = 0$, P is nonsingular, being given by

$$P = \int_0^\infty \exp(F't)LL' \exp(Ft) dt.$$

COROLLARY. The only matrices which commute with

$$(8) \quad \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}$$

are of the form

$$(9) \quad \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix},$$

where T_1, T_2' commute with F .

Proof of Corollary. Suppose that

$$\begin{bmatrix} T_1 & S \\ R & T_2 \end{bmatrix} \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix} = \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix} \begin{bmatrix} T_1 & S \\ R & T_2 \end{bmatrix}.$$

Then

$$(10a) \quad FT_1 = T_1F,$$

$$(10b) \quad F'T_2 = T_2F',$$

$$(10c) \quad FS + SF' = 0,$$

$$(10d) \quad RF + F'R = 0.$$

Clearly $S = 0$ and $R = 0$ satisfy (10c) and (10d). Lemma 5 guarantees these solutions are unique.

To conclude this section we state a lemma on spectral factorization, due to Youla [13]. Let $Z(s)$ be positive real. Let $Y(s) = Z(s) + Z'(-s)$; $Y(s)$ is termed *parahermitian*, i.e., $Y(s) = Y'(-s)$. If $s = j\omega$, $Y(j\omega)$ is non-negative definite hermitian as a consequence of the positive real character of $Z(s)$. Then [13, Theorem 2] yields a factorization of $Y(s)$ as follows.

LEMMA 6. Let the $n \times n$ matrix $Z(s)$ be positive real, and suppose that $Z(s) + Z'(-s)$ has rank r almost everywhere. Then there exists an $r \times n$ matrix $W(s)$ such that

$$(11) \quad Y(s) = Z(s) + Z'(-s) = W'(-s)W(s),$$

and

(i) W has elements which are analytic for $\operatorname{Re} s > 0$, and for $\operatorname{Re} s \geq 0$ if $Z(s)$ has elements which are analytic for $\operatorname{Re} s \geq 0$;

(ii) $\operatorname{rank} W = r$ for $\operatorname{Re} s > 0$;

(iii) W is unique save for multiplication on the left by an arbitrary orthogonal matrix.

3. Principal results. We shall assume until further notice that $Z(s)$ is positive real, with $Z(\infty) = 0$, and that it possesses no imaginary axis poles, i.e., all poles lie in the half-plane $\operatorname{Re} s < 0$. If $\{F, G, H\}$ is a minimal realization for Z , then F will have eigenvalues with negative real parts.

LEMMA 7. *Let $\{F, G, H\}$ be a minimal realization for $Z(s)$. With Z and W related as in Lemma 6, suppose W has a minimal realization $\{A, K, L\}$. Then the matrices A and F are similar.*

Proof. Because $\{F, G, H\}$ is a realization for $Z(s)$, direct calculation shows that

$$(12) \quad \{F_1, G_1, H_1\} = \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ H \end{bmatrix}, \begin{bmatrix} H \\ -G \end{bmatrix} \right\}$$

is a realization for $Z'(-s) + Z(s)$. Because $Z(s)$ and $Z'(-s)$ can have no poles in common (those of $Z(s)$ being in $\operatorname{Re} s < 0$ and those of $Z'(-s)$ in $\operatorname{Re} s > 0$), by Lemma 1,

$$\delta[Z(s) + Z'(-s)] = 2\delta[Z(s)].$$

Hence $\{F_1, G_1, H_1\}$ is minimal.

By direct calculation

$$(13a) \quad W'(-s)W(s) = K'(-sI - A')^{-1}LL'(sI - A)^{-1}K$$

$$(13b) \quad = H_2'(sI - F_2)^{-1}G_2,$$

where

$$(14) \quad \{F_2, G_2, H_2\} = \left\{ \begin{bmatrix} A & 0 \\ LL' & -A' \end{bmatrix}, \begin{bmatrix} K \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -K \end{bmatrix} \right\}.$$

By Lemma 4 and (i) and (ii) in Lemma 6, $\delta[W'(-s)W(s)] = 2\delta[W]$, and thus $\{F_2, G_2, H_2\}$ is minimal.

Define P to be the unique positive definite symmetric solution of

$$PA + A'P = -LL'.$$

The existence of P follows by Lemma 5, (i) in Lemma 6 and the minimality of $\{A, K, L\}$. We may now apply Lemma 1, taking

$$T = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix}$$

to obtain the following alternative minimal realization of $W'(-s)W(s)$:

$$(15) \quad \{F_3, G_3, H_3\} = \left\{ \begin{bmatrix} A & 0 \\ 0 & -A' \end{bmatrix}, \begin{bmatrix} -K \\ PK \end{bmatrix}, \begin{bmatrix} PK \\ -K \end{bmatrix} \right\}.$$

Since (12) and (15) are minimal realizations of the same matrix, viz., $Y(s)$, and since F has strictly negative eigenvalues, as also has A , from (i) in Lemma 6, the result of the lemma follows by (3a) in Lemma 1.

COROLLARY A. Let $Z(s)$ have a minimal realization $\{F, G, H\}$ and let Z and W be related as in Lemma 6. Then there exist matrices K, L such that W has a minimal realization $\{F, K, L\}$.

COROLLARY B. With the notation of Corollary A, two minimal realizations of $Y(s) = Z(s) + Z'(-s) = W'(-s)W(s)$ are given by

$$(12) \quad \{F_1, G_1, H_1\} = \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ H \end{bmatrix}, \begin{bmatrix} H \\ -G \end{bmatrix} \right\}$$

and

$$(16) \quad \{F_3, G_3, H_3\} = \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} P \\ PK \end{bmatrix}, \begin{bmatrix} PK \\ -K \end{bmatrix} \right\},$$

where P is uniquely defined (see Lemma 5) by

$$(17) \quad PF + F'P = -LL',$$

and K, L are now as defined in Corollary A.

The proofs of these two corollaries follow from Lemma 1.

The next lemma is concerned with making a further identification between minimal realizations of Z and W .

LEMMA 8. Let $Z(s)$ have a minimal realization $\{F, G, H\}$ and let Z and W be related as in Lemma 6. Then there exists a matrix \hat{L} such that $\{F, G, \hat{L}\}$ is a minimal realization for W .

Proof. By Corollary A of the previous lemma, W has a minimal realization $\{F, K, L\}$ for some K and L . By Corollary B and Lemma 1 there exists a nonsingular matrix T which must commute with

$$\begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}$$

such that

$$T \begin{bmatrix} G \\ H \end{bmatrix} = \begin{bmatrix} K \\ PK \end{bmatrix}$$

By Lemma 5, there exists T_1 commuting with F such that

$$T_1 G = K.$$

Moreover, since T is nonsingular, so is T_1 . Then by Lemma 1 again, since $\{F, K, L\}$ is a minimal realization for W , so is $\{T_1^{-1}FT_1, T_1^{-1}K, (T_1)'L\} = \{F, G, \hat{L}\}$.

We can now state the following theorem.

THEOREM 1. Let $Z(\cdot)$ be a matrix of rational functions such that $Z(\infty) = 0$ and Z has poles only in $\text{Re } s < 0$. Let $\{F, G, H\}$ be a minimal realization of Z . Then $Z(\cdot)$ is positive real if and only if there exist a symmetric positive defi-

nite matrix P and a matrix L such that

$$(17) \quad PF + F'P = -LL'$$

and

$$(18) \quad PG = H.$$

Proof of sufficiency. Of the three conditions listed in §1 which Z must satisfy, the only one which needs to be verified is the third, viz., $Z'(s^*) + Z(s)$ is nonnegative definite for $\operatorname{Re} s > 0$. We have

$$\begin{aligned} Z'(s^*) + Z(s) &= G'(s^*I - F')^{-1}H + H'(sI - F)^{-1}G \\ &= G'\{(s^*I - F')^{-1}P + P(sI - F)^{-1}\}G \\ &= G'(s^*I - F')^{-1}\{P(sI - F) + (s^*I - F')P\}(sI - F)^{-1}G \\ &= G'(s^*I - F')^{-1}\{Ps + s^*P - PF - F'P\}(sI - F)^{-1}G \\ &= G'(s^*I - F')^{-1}P(sI - F)^{-1}G(s + s^*) \\ &\quad + G'(s^*I - F')^{-1}LL'(sI - F)^{-1}G. \end{aligned}$$

Since the right-hand side is of the form $\{2\operatorname{Re} s\}\{A'(s^*)PA(s) + B'(s^*)B(s)\}$, it is clearly nonnegative definite in the right half-plane.

Proof of necessity. Let $W(s)$ be as in Lemma 6, and let $\{F, G, L\}$ be a minimal realization of W , where we are using Lemma 8 and dropping the hat notation. The matrix L is used to define through (17) the matrix P which we know from Lemma 5 to be unique and symmetric positive definite. By Corollary B to Lemma 7,

$$\left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ H \end{bmatrix}, \begin{bmatrix} H \\ -G \end{bmatrix} \right\} \text{ and } \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ PG \end{bmatrix}, \begin{bmatrix} PG \\ -G \end{bmatrix} \right\}$$

are two minimal realizations of $Y(s)$. By Lemma 1, there exists nonsingular T commuting with

$$\begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}$$

such that

$$T \begin{bmatrix} G \\ H \end{bmatrix} = \begin{bmatrix} G \\ PG \end{bmatrix} \quad \text{and} \quad (T^{-1})' \begin{bmatrix} H \\ -G \end{bmatrix} = \begin{bmatrix} PG \\ -G \end{bmatrix}.$$

By Lemma 5, there exists T_1 commuting with F such that $T_1G = G$ and $(T_1^{-1})'H = PG$. Now, since T_1 commutes with F ,

$$[G, FG, \dots] = [T_1G, FT_1G, \dots] = [T_1G, T_1FG, \dots] = T_1[G, FG, \dots].$$

The matrix $[G, FG, \dots]$ has rank p , where p is the dimension of F , since $\{F, G, H\}$ is a minimal realization of $Z(s)$ and all minimal realizations are completely controllable [5]. Hence $T_1 = I$, and thus $PG = H$.

In preparation for dealing with matrices which may possess imaginary axis poles, we consider now positive real matrices whose *only* poles are on the imaginary axis.

THEOREM 2. *Let a positive real $Z(s)$ have all pure imaginary poles, with $Z(\infty) = 0$, and let $\{F, G, H\}$ be a minimal realization for Z . Then there exists a symmetric positive definite P such that*

$$(19) \quad PF + F'P = 0,$$

$$(20) \quad PG = H.$$

Proof. First note that if P satisfies the above equations, then $P^* = (T')^{-1}PT^{-1}$ satisfies the corresponding equations for the minimal realization $\{TFT^{-1}, TG, (T^{-1})'H\}$. Consequently if we exhibit a symmetric positive definite P for any one minimal realization of Z , it follows that a symmetric positive definite P exists for all minimal realizations. Our procedure will in fact be to choose a minimal realization $\{F, G, H\}$ for which P has a particularly obvious form.

The form of $Z(s)$ has been established (see, for example, [2]) as

$$(21) \quad Z(s) = \sum_i \frac{A_i s + B_i}{s^2 + \omega_i^2},$$

where the ω_i are all different and the matrices A_i and B_i satisfy certain requirements. By realizing separately each term $(A_i s + B_i)(s^2 + \omega_i^2)^{-1}$ with minimal $\{F_i, G_i, H_i\}$ and selecting a P_i such that (19) and (20) are satisfied, one can obtain a minimal $\{F, G, H\}$ and a P satisfying (19) and (20) with $F = \dot{+}_i F_i$ (where $\dot{+}$ denotes direct sum), $G' = [G_1', G_2', \dots]$, $H' = [H_1', H_2', \dots]$ and $P = \dot{+}_i P_i$. Consequently we shall consider the realization of the simpler

$$(22) \quad Z(s) = \frac{As + B}{s^2 + \omega_0^2}.$$

In [2, Chap. 6] it is pointed out that if $2k$ is the degree of Z in (22), there exist k complex vectors x_i such that

$$(23) \quad x_i^{*'} x_i = 1, \quad x_i' x_i = \mu_i, \quad 0 < \mu_i \leq 1, \quad \mu_i \text{ real},$$

and

$$(24) \quad Z(s) = \sum_{i=1}^k \left[\frac{x_i x_i^{*'}}{s - j\omega_0} + \frac{x_i^* x_i'}{s + j\omega_0} \right].$$

Direct sum techniques then allow us to restrict consideration to obtaining

a minimal realization for the degree 2

$$(25) \quad Z(s) = \frac{xx^{*'}}{s - j\omega_0} + \frac{x^*x'}{s + j\omega_0}.$$

It is easy now to verify that if

$$(26) \quad y_1 = \frac{x + x^*}{\sqrt{2}}, \quad y_2 = j \frac{x - x^*}{\sqrt{2}},$$

we have

$$(27) \quad Z(s) = [y_1, y_2] \frac{1}{s^2 + \omega_0^2} \begin{bmatrix} s & \omega_0 \\ -\omega_0 & s \end{bmatrix} \begin{bmatrix} y_1' \\ y_2' \end{bmatrix},$$

and then

$$(28) \quad \{F, G, H, P\} = \left\{ \begin{bmatrix} 0 & -\omega_0 \\ \omega_0 & 0 \end{bmatrix}, \begin{bmatrix} y_1' \\ y_2' \end{bmatrix}, \begin{bmatrix} y_1' \\ y_2' \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

defines a minimal realization for (25), with (19) and (20) being satisfied.

All preliminaries are now in hand to give the final theorem, which applies to general positive real matrices. Since any positive real matrix can be written as the sum of D s and $Z(s)$, where D is nonnegative definite and $Z(s)$ is also positive real, but with $Z(\infty)$ finite, we shall lose no generality in restricting the theorem statement to such $Z(s)$ (see [2, Chap. 5]).

By way of notation, we shall say that $\{F, G, H, Z(\infty)\}$ is a minimal realization of $Z(s)$ if $\{F, G, H\}$ is a minimal realization of $Z(s) - Z(\infty)$.

THEOREM 3. *Let $Z(\cdot)$ be a matrix of rational transfer functions such that $Z(\infty)$ is finite and Z has poles which lie in $\text{Re } s < 0$ or are simple on $\text{Re } s = 0$. Let $\{F, G, H, Z(\infty)\}$ be a minimal realization of Z . Then $Z(\cdot)$ is positive real if and only if there exist a symmetric positive definite P and matrices W_0 and L such that*

$$(29) \quad PF + F'P = -LL',$$

$$(30) \quad PG = H - LW_0,$$

$$(31) \quad W_0'W_0 = Z(\infty) + Z'(\infty).$$

Proof of sufficiency. It only remains to verify the positive real behavior of $Z'(s^*) + Z(s)$ in the right half-plane. We have

$$\begin{aligned} Z'(s^*) + Z(s) &= Z'(\infty) + Z(\infty) + G'(s^*I - F')^{-1}H + H'(sI - F)^{-1}G \\ &= W_0'W_0 + G'[(s^*I - F')^{-1}P + P(sI - F)^{-1}]G \\ &\quad + G'(s^*I - F')^{-1}LW_0 + W_0'L'(sI - F)^{-1}G \end{aligned}$$

$$\begin{aligned}
(32) \quad &= W_0' W_0 + G'(s^* I - F')^{-1} [P(s + s^*) - PF - F'P] (sI - F)^{-1} G \\
&\quad + G'(s^* I - F')^{-1} L W_0 + W_0' L' (sI - F)^{-1} G \\
&= W_0' W_0 + G'(s^* I - F')^{-1} L W_0 + W_0' L' (sI - F)^{-1} G \\
&\quad + G'(s^* I - F')^{-1} L L' (sI - F)^{-1} G \\
&\quad + G'(s^* I - F')^{-1} P (sI - F)^{-1} G (s + s^*) \\
&= [W_0' + G'(s^* I - F')^{-1} L] [W_0 + L' (sI - F)^{-1} G] \\
&\quad + G'(s^* I - F')^{-1} P (sI - F)^{-1} G (s + s^*),
\end{aligned}$$

which is plainly nonnegative definite for $\text{Re } s > 0$.

This completes the proof of sufficiency.

Proof of necessity. Initially, suppose $Z(s)$ has strict left half-plane poles. We shall consider the general case later, with the aid of Theorem 2. Let $W(s)$ be the matrix defined in Lemma 6. Then the arguments used to establish Lemmas 7 and 8 carry through in essentially the same fashion to establish that there exist matrices L and $W_0 = W(\infty)$ such that W has a minimal realization $\{F, G, L, W_0\}$, with two minimal realizations for $Z(s) + Z'(-s) = W'(-s)W(s)$ being given by

$$(33) \quad \{F_1, G_1, H_1, W_0' W_0\} = \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ H \end{bmatrix}, \begin{bmatrix} H \\ -G \end{bmatrix}, W_0' W_0 \right\}$$

and

$$\begin{aligned}
(34) \quad &\{F_3, G_3, H_3, W_0' W_0\} \\
&= \left\{ \begin{bmatrix} F & 0 \\ 0 & -F' \end{bmatrix}, \begin{bmatrix} G \\ PG + LW_0 \end{bmatrix}, \begin{bmatrix} PG + LW_0 \\ -G \end{bmatrix}, W_0' W_0 \right\}.
\end{aligned}$$

Here P is the unique symmetric positive definite solution of (29).

The arguments of Theorem 1 can now be followed directly to conclude that (30) holds. Equation (31) follows by setting $s = \infty$ in $Z(s) + Z'(-s) = W'(-s)W(s)$.

Now let us relax the restriction on the poles of $Z(s)$. Then $Z(s)$ may be decomposed as

$$(35) \quad Z(s) = Z_1(s) + Z_2(s),$$

where Z_1 has purely imaginary axis poles, Z_2 has poles only in $\text{Re } s < 0$, and Z_1 and Z_2 are both positive real [2, §5.2].

Now select for Z_1 a minimal $\{F_1, G_1, H_1\}$ and P_1 , using Theorem 2, such that

$$(19') \quad P_1 F_1 + F_1' P_1 = 0,$$

$$(20') \quad P_1 G_1 = H_1 ;$$

and for Z_2 select $\{F_2, G_2, H_2\}$ and P_2 , using the material just proved, such that

$$(29') \quad P_2 F_2 + F_2' P_2 = -L_2 L_2',$$

$$(30') \quad P_2 G_2 = H_2 - L_2 W_0,$$

$$(31') \quad W_0' W_0 = Z_2(\infty) + Z_2'(\infty).$$

Then it is easily verified that (29), (30) and (31) are satisfied by taking

$$(36) \quad \begin{aligned} P &= P_1 \dot{+} P_2, \\ F &= F_1 \dot{+} F_2, \\ G' &= [G_1', G_2'], \\ H' &= [H_1', H_2'], \\ L' &= [0, L_2']. \end{aligned}$$

Moreover, with $\{F_1, G_1, H_1\}$ and $\{F_2, G_2, H_2\}$ minimal realizations for $Z_1(s)$ and $Z_2(s) - Z_2(\infty)$, $\{F, G, H\}$ is a minimal realization for $Z(s) - Z(\infty)$. This is because, by Lemma 2, the degree of Z is the sum of the degrees of Z_1 and Z_2 , the latter having no common poles, while the dimension of F is the sum of the dimensions of F_1 and F_2 . One should at this stage verify that (29), (30) and (31) are valid under a state space coordinate transformation, as they have merely been established for a particular class of F (i.e., those of the form $F_1 \dot{+} F_2$). This is easy to do, however, along the lines given in Theorem 2 for a more particular case.

4. Conclusions. The significance of the theorems in their own right is self-evident. They provide a conceptual link between basic concepts of control theory and network theory. Their proofs have a number of interesting features, such as the necessity to use the particular $W(s)$ in the factorization of $Z(s) + Z'(-s)$, the heavy reliance on the concept of degree in the network [8] and control theory [11] senses, and the canonical $\{F, G, H\}$ representation (believed new) of a "lossless" Z . Hopefully the results themselves as well as their proofs will help forge another link in the growing chain [3], [14] between control and network theory.

There are several immediate applications of the theory. The stability of control systems containing multiple nonlinearities is discussed in [15], a new passive network synthesis of positive real functions and matrices in [16], and the properties of multivariable control systems with linear feedback laws in [17].

REFERENCES

- [1] R. E. KALMAN, *Lyapunov functions for the problem of Lur'e in automatic control*, Proc. Nat. Acad. Sci. U. S. A., 49 (1963), pp. 201-205.
- [2] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.
- [3] R. E. KALMAN, *On a new characterization of linear passive systems*, Proc. First Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1963, pp. 456-470.
- [4] V. M. POPOV, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Electrotechn et Energ., 9 (1964), pp. 629-690.
- [5] R. E. KALMAN, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.
- [6] B. D. O. ANDERSON AND R. W. NEWCOMB, *A canonical simulation of a transfer function matrix*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 319-320.
- [7] B. L. HO AND R. E. KALMAN, *Effective construction of linear state-variable models from input/output data*, Proc. Third Allerton Conference on Circuit and System Theory, University of Illinois, Urbana, 1965.
- [8] B. McMILLAN, *Introduction to formal realizability theory*, Bell System Tech. J., 31 (1952), pp. 217-279; 31 (1952), pp. 541-600.
- [9] R. J. DUFFIN AND D. HAZONY, *The degree of a rational matrix function*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 645-658.
- [10] D. HAZONY, *Elements of Network Synthesis*, Reinhold, New York, 1963.
- [11] R. E. KALMAN, *Irreducible realizations and the degree of a matrix of rational functions*, J. Soc. Indust. Appl. Math., 13 (1965), pp. 520-544.
- [12] J. LaSALLE AND S. LEFSCHETZ, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.
- [13] D. C. YOULA, *On the factorization of rational matrices*, IRE Trans. Information Theory, IT-7 (1961), pp. 172-189.
- [14] D. C. YOULA AND P. TISSI, *N-port synthesis via reactance extraction, Part I*, Electrophysics Memo PIBMRI-1309-66, Polytechnic Institute of Brooklyn, Brooklyn, New York, 1966.
- [15] B. D. O. ANDERSON, *Stability of control systems with multiple nonlinearities*, J. Franklin Inst., 281 (1966), pp. 155-160.
- [16] B. D. O. ANDERSON AND R. W. NEWCOMB, *Impedance synthesis via state space techniques*, Rept. SEL-66-024 (TR 6558-5), Stanford Electronics Laboratories, Stanford, California, 1966.
- [17] B. D. O. ANDERSON, *The inverse problem of optimal control*, Rept. SEL-66-038 (TR 6560-3), Stanford Electronics Laboratories, Stanford, California, 1966.

CONTROLLABILITY OF DISCRETE, LINEAR, RANDOM DYNAMICAL SYSTEMS*

MICHAEL M. CONNORS†

1. Introduction. This paper is concerned with a detailed study of the concept of controllability in discrete, linear, random dynamical systems. Despite increasing interest in stochastic control problems, little interest has been shown in the fundamental question of stochastic controllability. The very first question which arises in a rigorous theory of stochastic control is: "Can an initial state of a given random dynamical process be transferred (in a stochastic sense) to any desired state in a finite length of time by some control action?" In analogy with the theory of deterministic controllability (cf. [11]), if the answer is in the affirmative, we shall say that the initial state is stochastically controllable.

This paper will formulate a precise definition of stochastic controllability and will give a detailed analysis of the problem for a general class of linear, random dynamical systems.

In addition to mathematical characterizations of stochastic controllability, the analysis will yield a number of qualitative results relating to linear random dynamical processes. The application of the concept of stochastic controllability to particular stochastic optimal control problems is discussed, and the extension of the idea of stochastic controllability to linear adaptive systems is analyzed.

2. Problem statement and preliminaries. We shall use the standard notation and terminology. Lower case letters a, b, c and u, v, \dots, z will denote column vectors with (real) components a_i, b_i , etc. Superscripts on vectors are used to distinguish them from other vectors of the same genre. The letters h, i, j, k, \dots, s, t denote integers. The greek letters $\epsilon, \tau, \sigma, \rho$ denote scalars. The capital letters A, B, \dots, F, G and P, Q, \dots, S, T denote matrices with components a_{ij}, b_{ij} , etc. The columns of these matrices are α^i, β^i , etc., and the rows are α_i, β_i , etc. The capital letters U, V, \dots, Z denote collections of vectors, linear spaces, etc. The capital letters I, J, \dots, M, N are reserved for constants. Time is always denoted by the integers h, i, j, k, \dots, s, t ; constant values of time are ordered in the obvious manner: time $i < \text{time } j$ if $i < j$. The transpose of a matrix is denoted by a prime. The Euclidean inner product is the only inner product used and is denoted by $\langle x, y \rangle$ or by $x \cdot y$. The norm corresponding to the Euclidean inner

* Received by the editors August 8, 1966, and in revised form November 11, 1966.

† Program in Operations Research, Stanford University, Stanford, California.
Now at IBM Scientific Center, Los Angeles, California.

product is always written $\|x\| = \langle x, x \rangle^{1/2}$. If A is a symmetric, nonnegative definite matrix, we write $\langle x, Ax \rangle = \|x\|_A^2$ to denote the generalized Euclidean norm and $\langle x, Ay \rangle = \langle x, y \rangle_A$. The eigenvalues of a matrix A are denoted by $\lambda_i[A]$ unless the meaning is apparent from the context in which case we write λ_i . The identity matrix is I . Certain algebraic quantities such as algebras, fields, Lie algebras, etc., are denoted by script capitals $\mathfrak{A}, \dots, \mathfrak{F}, \dots, \mathfrak{L}, \dots$. Occasionally, certain special spaces or collections will also be denoted by script capitals. Mathematical expectation is denoted by \mathfrak{E} .

We shall consider the particular system which can be described by the stochastic vector difference equation

$$(1) \quad x(i+1) = A(i, \omega_i)x(i) + B(i, \omega_i)u(i, x(i)), \quad x(0) = x^0,$$

where $x(i)$ is the n -vector which specifies the state of the process at time i ; $u(i, x(i))$ is the m -vector which represents the control action taken at time i ; ω_i is the i th component of the sequence $\{\omega_i\}_{i=0}^\infty$ which is a sequence of independent random variables with distribution functions $F_i, i = 0, 1, \dots$; $A(i, \omega_i)$ is an $n \times n$ (random) matrix which may vary over time and which depends on the random variable ω_i (that is, $A(i, \omega_i) = (a_{st}(i, \omega_i))$); $B(i, \omega_i)$ is an $n \times m$ (random) matrix defined in the same manner as $A(i, \omega_i)$; $x(0) = x^0$ is the initial condition on the dynamical process; $i = 0, 1, \dots$ is the index whose value specifies the time period under consideration.

It will be convenient to denote the state of the process (1) at the terminal time k by $x(k)$ or, equivalently, as $\phi(k; x^0, 0, u_{[0,k]})$ (read: "the trajectory of the process evaluated at the terminal sampling period k given the initial condition x^0 at sampling period 0 and given that the process has evolved under influence of the control law $u_{[0,k]}$ "). The control law, $u_{[0,k]}$, is the sequence of m -vectors $u(i, x(i)), i = 0, 1, \dots, k-1$, where each $u(i, x(i))$ is defined above. The function $\phi(k; x^0, 0, u_{[0,k]})$ is the solution of system (1). In other words, the function $\phi(k; x^0, 0, u_{[0,k]})$ has the following properties:

- (i) it satisfies the initial condition $\phi(0; x^0, 0, u_{[0,k]}) = x^0$;
- (ii) it satisfies (1) for all $i = 0, 1, \dots, k$;
- (iii) it is a random function depending on the random variables $\omega_i, i = 0, 1, \dots, k-1$, and has the same distribution as any other random function which satisfies (1).

If $u_{[0,k]}$ is the sequence of zero m -vectors, we say that the motion is free and denote it by $\phi(k; x^0, 0, 0)$. The notation $x(k)$ will usually be used to denote the state of the process at time k ; the (equivalent) notation $\phi(k; x^0, 0, u_{[0,k]})$ will be used when it is desirable to indicate the dependence of the state of the process at sampling period k upon the initial conditions and the control law.

It is easy to obtain an explicit formula for $x(k) = \phi(k; x^0, 0, u_{[0,k]})$. This can be given as

$$(2) \quad \begin{aligned} x(k) &= \phi(k; x^0, 0, u_{[0,k]}) \\ &= \Phi(k, 0)x^0 + \sum_{i=1}^k \Phi(k, i)B(i-1, \omega_{i-1})u(i-1, x(i-1)), \end{aligned}$$

where

$$\begin{aligned} \Phi(k, i) &= A(k-1, \omega_{k-1})A(k-2, \omega_{k-2}) \cdots A(i, \omega_i), \\ \Phi(i, i) &= I. \end{aligned}$$

The particular problem in the theory of control to be studied is that of stochastic controllability of system (1). We shall need the following definition.

DEFINITION 1. An initial state x^0 is *ϵ -controllable in norm square* (abbreviated *ϵ -c.n.s.*) with respect to a specified terminal state, x^f , in the time interval $[0, k]$ if and only if there exists a control law

$$u_{[0,k]} = \{u(i, x(i)), i = 0, \dots, k-1\}$$

such that

$$(3) \quad \mathcal{E}[\|x^f - \phi(k; x^0, 0, u_{[0,k]})\|^2] \leq \epsilon(\|x^0\|^2 + \|x^f\|^2),$$

where the norm is the Euclidean norm and \mathcal{E} is the expected value operator taken with respect to the joint probability distribution function of the independent random variables ω_i , $i = 0, 1, \dots, k-1$.

DEFINITION 2. The process (1) is *completely ϵ -controllable in norm square* with respect to a specified terminal state, x^f , in the time interval $[0, k]$ if and only if inequality (3) holds for every initial state $x^0 \in E^n$.

Remark. It is of interest to note that Definitions 1 and 2 can be given in terms of a generalized Euclidean norm instead of the standard Euclidean norm. That is, inequality (3) may be written as

$$\mathcal{E}[\|x^f - \phi(k; x^0, 0, u_{[0,k]})\|_D^2] \geq \epsilon(\|x^0\|_F^2 + \|x^f\|_G^2),$$

where D , F and G are symmetric nonnegative definite matrices and $\|x\|_D^2 = x'Dx$ and similarly for F and G . If it were desired to give various weights to different components of the difference between x^f and $x(k)$ at the terminal sampling period, we would define D accordingly. Similarly weighting matrices F and G could be constructed to give various weights to components of the initial state x^0 or the terminal state x^f . Note also that D , F and G may be defined independently of each other. These cases may be treated in exactly the same manner as the case $D = F = G = I$ which is discussed here. It is important to note that the use of the Euclidean norm

(in either its generalized or standard form) is necessary for the development of the theory in the present paper.

Remark. A similar problem, involving stochastic differential equations with additive random noise, has been treated by Krasovskii [13]. In [13], the terminal criterion is a probabilistic rather than norm square criterion as treated here. It is of interest to note that the results in [13] are obtained by solving a norm square terminal criterion problem and then reducing the result to a probabilistic terminal criterion via the Chebyshev inequality.

A loose verbalization of the preceding definitions is that an initial state (or the process (1)) is ϵ -c.n.s. if and only if there exists a control law which drives the "stochastic distance" between x^f and the state of the system at time k to some ϵ quantity.

3. Mathematical characterization of controllability. In this section, dynamic programming will be used to yield necessary and sufficient conditions for the process (1) to be stochastically controllable. In order to develop the characterization in greatest generality, it will be necessary to use the concept of the generalized inverse of a matrix introduced by Penrose [15], [16]. We shall need a third definition.

DEFINITION 3. The generalized inverse R^\dagger of an arbitrary real rectangular matrix R is a matrix satisfying the relations:

- (i) $RR^\dagger R = R$,
- (ii) $R^\dagger RR^\dagger = R^\dagger$,
- (iii) $(R^\dagger R)' = R^\dagger R$,
- (iv) $(RR^\dagger)' = RR^\dagger$.

Hence, $R^\dagger = R^{-1}$ when R is invertible. Note also that $R^{\dagger\dagger} = R$ and that $(R')^\dagger = (R^\dagger)'$. Relations (i)–(iv) imply that R^\dagger always exists and is unique. It will be useful to note that:

- (1) if D is diagonal, the elements of D^\dagger are

$$d_{ii}^\dagger = \begin{cases} d_{ii}^{-1} & \text{if } d_{ii} \neq 0, \\ 0 & \text{otherwise;} \end{cases}$$

- (2) if S is symmetric, there is an orthogonal transformation Γ such that $S = \Gamma D \Gamma'$; then $S^\dagger = \Gamma' D^\dagger \Gamma$.

The proofs of statements (1) and (2) may be found in Kalman [12].

Returning to the problem of controllability, it is apparent that the condition of Definition 1 will be satisfied if and only if there exists a control law $u_{[0,k]}$ such that

$$\inf_{u_{[0,k]}} \mathbb{E}[\|x^f - \phi(k; x^0, 0, u_{[0,k]})\|^2] \leq \epsilon(\|x^0\|^2 + \|x^f\|^2).$$

It is now possible to employ the dynamic programming technique and the concept of the pseudoinverse of a matrix to calculate a control law $u_{[0,k]}$ which minimizes $\mathbb{E}[\|x^f - \phi(k; x^0, 0, u_{[0,k]})\|^2]$.

In order to avoid details, the following work will be based on the special case $x^f = 0$. It will be apparent that similar results may be obtained in the general case; moreover, the same techniques are used in the proofs for the more general case. Focusing attention on the special case $x^f = 0$ is justified by noting that any problem can be reduced to this case by a translation of coordinates such that, in the new coordinate system, x^f is the origin.

Using the notation $A_i = A(i, \omega_i)$ and $B_i = B(i, \omega_i)$ the result can be stated as the following lemma.

LEMMA 1. *Let $x^f = 0$. Then the control law $u_{[0,k]}^* = \{u^*(i, x(i)), i = 0, \dots, k-1\}$ which minimizes $\mathbb{E}[\|\phi(k; x^0, 0, u_{[0,k]})\|^2]$ is given by*

$$u^*(i, x(i)) = -\mathbb{E}_i[B_i'Q(k-i-1)B_i]^\dagger \mathbb{E}_i[B_i'Q(k-i-1)A_i]x(i),$$

where

$$\begin{aligned} Q(k-i) &= \mathbb{E}_i[A_i'Q(k-i-1)A_i] \\ &- \mathbb{E}_i[A_i'Q(k-i-1)B_i] \mathbb{E}_i[B_i'Q(k-i-1)B_i]^\dagger \mathbb{E}_i[B_i'Q(k-i-1)A_i], \\ Q(0) &= I, \end{aligned}$$

and where the symbol \mathbb{E}_i denotes the expected value operator taken with respect to the probability distribution of the random variable ω_i , $i = 0, 1, \dots, k-1$.

Moreover, the value of the performance index, when there are $k-i$ stages left in the process and the present state is $x(i)$, is

$$f_{k-i}(x(i)) = \|x(i)\|_{Q(k-i)}^2.$$

Proof. Since the random variables ω_j are independent, knowledge of $x(j-1)$, $x(j-2)$, \dots conveys no additional information about the future evolution of (1). This implies that knowledge of the present state vector together with knowledge of the distribution function of ω constitutes a complete state description at sampling period j (i.e., the system is Markovian). Hence, the Principle of Optimality may be applied. Thus, we define

$$\begin{aligned} f_{k-i}(x) &= \text{minimum value of } \mathbb{E}[\|x(k)\|^2] \text{ over the } k-i \text{ remaining sam-} \\ &\quad \text{pling periods of the process, starting in state } x \text{ at sampling} \\ &\quad \text{period } i, \text{ subject to (1) and using an optimal policy for the} \\ &\quad \text{remaining sampling periods, where } \mathbb{E} \text{ now denotes the ex-} \\ &\quad \text{pected value operator taken with respect to the joint} \\ &\quad \text{probability distribution function of the random variables} \\ &\quad \omega_j, j = i, i+1, \dots, k-1, \\ &= \min_{u_{[i,k]}} \mathbb{E}[\|x(k)\|^2] \\ &= \min_{u(i,x)} \mathbb{E}_i[f_{k-i-1}(A(i, \omega_i)x + B(i, \omega_i)u(i, x))]. \end{aligned}$$

It is assumed by induction that $f_{k-i-1}(x) = x'Q(k-i-1)x$, where $Q(k-i)$ is defined in the statement of the lemma. It should be noted that the first step in the induction is trivial and is identical to the general step

and hence will not be given. Then, using a technique suggested by Kalman [9], let

$$S(i) = \varepsilon_i[A_i'Q(k-i-1)A_i],$$

$$T(i) = \varepsilon_i[B_i'Q(k-i-1)A_i],$$

$$U(i) = \varepsilon_i[B_i'Q(k-i-1)B_i],$$

and write

$$\begin{aligned} f_{k-i}(x) &= \min_{u(i,x)} \varepsilon_i(f_{k-i-1}(A_ix + B_iu(i,x))) \\ &= \min_{u(i,x)} \varepsilon_i[\|A_ix + B_iu(i,x)\|_{Q(k-i-1)}^2] \\ (4) \quad &= \min_{u(i,x)} \|x\|_{S(i)-T'(i)U^\dagger(i)T(i)}^2 + \|T(i)x + U(i)u(i,x)\|_{U^\dagger(i)}^2 \\ &\quad + 2\langle T(i)x, (I - U^\dagger(i)U(i))u(i,x) \rangle, \end{aligned}$$

where use has been made of relation (i) in Definition 3. The term in the fourth line of (4) is zero since, if it were equal to $\alpha \neq 0$ for, say, x^1, u^1 , then substituting

$$u(i,x) = \alpha\lambda(I - U^\dagger(i)U(i))u^1$$

shows that, for any λ ,

$$f_{k-i}(x^1) \leq \|x^1\|_{S(i)}^2 + 2\alpha^2\lambda,$$

which is absurd since $f_{k-i}(x) \geq 0$ by definition. This contradiction establishes that $\langle T(i)x, (I - U^\dagger(i)U(i))u(i,x) \rangle = 0$.

Thus, the minimum of $f_{k-i}(x)$ is attained when the term

$$\|T(i)x + U(i)u(i,x)\|_{U^\dagger(i)}^2$$

equals zero. Letting

$$u^*(i,x) = -U^\dagger(i)T(i)x$$

shows that

$$\begin{aligned} \|T(i)x + U(i)u^*(i,x)\|_{U^\dagger(i)}^2 &= \|(I - U(i)U^\dagger(i))T(i)x\|_{U^\dagger(i)}^2 \\ &= \|(I - U^\dagger(i)U(i))T(i)x\|_{U^\dagger(i)}^2 \\ &= 0, \end{aligned}$$

where use has been made firstly of the fact that $U(i)$ is symmetric and, hence, so is $U^\dagger(i)$ and secondly, of relation (iv) of Definition 3.

Thus the minimum of $f_{k-i}(x)$ is assumed for $u^*(i,x)$ which verifies the

first statement of the lemma. Also

$$\begin{aligned} f_{i-1}(k) &= \|x\|_{S(i)-T'(i)U^\dagger(i)T(i)}^2 \\ &= \|x\|_{Q(k-i)}^2, \end{aligned}$$

which completes the proof.

We state the general result for completeness. The proof is accomplished using the technique of the proof of Lemma 1.

LEMMA 2. *The control law $u_{[0,k]}^* = \{u^*(i, x(i)), i = 0, 1, \dots, k-1\}$ which minimizes $\mathcal{E}[\|x^f - \phi(k; x^0, 0, u_{[0,k]})\|^2]$ is given by*

$$\begin{aligned} u^*(i) &= -\varepsilon_i[B_i'Q(k-i-1)B_i]^\dagger \{\varepsilon_i[B_i'Q(k-i-1)A_i]x(i) \\ &\quad - \varepsilon_i[B_i']P(k-i-1)x'\}, \end{aligned}$$

where ε_i and $Q(k-i)$ are defined in Lemma 1 and

$$\begin{aligned} P(k-i) &= \varepsilon_i[A_i' - \varepsilon_i[A_i'Q(k-i-1)B_i] \\ &\quad \cdot \varepsilon_i[B_i'Q(k-i-1)B_i]^\dagger B_i']P(k-i-1), \\ P(0) &= I. \end{aligned}$$

Moreover, the value of the performance index, when there are $k-i$ stages left in the process and the present state is $x(i)$, is

$$f_{k-1}(x(i)) = \|x(i)\|_{Q(k-i)}^2 - 2x'(i)P(k-i)x^f + \|x^f\|_{S(k-i)}^2,$$

where

$$\begin{aligned} S(k-i) &= S(k-i-1) \\ &\quad + P'(k-i-1)\varepsilon_i[B_i]\varepsilon_i[B_i'Q(k-i-1)B_i]^\dagger P(k-i-1). \\ S(0) &= I. \end{aligned}$$

Using the control law given in Lemma 1, the system (1) can be written

$$\begin{aligned} x(i+1) &= \{A_i - B_i\varepsilon_i[B_i'Q(k-i-1)B_i]^\dagger \varepsilon_i[B_i'Q(k-i-1)A_i]\}x(i) \\ &= \Psi(i)x(i), \end{aligned}$$

where $\Psi(i)$ is the expression in curly brackets. This leads to a corollary.

COROLLARY 1.

$$Q(k-i) = \varepsilon_i[\Psi'(i)Q(k-i-1)\Psi(i)] \quad \text{for } i = 0, 1, \dots, k-1,$$

where

$$\Psi(i) = A_i - B_i\varepsilon_i[B_i'Q(k-i-1)B_i]^\dagger \varepsilon_i[B_i'Q(k-i-1)A_i].$$

Proof. By direct substitution.

COROLLARY 2. $Q(k-i)$ is positive semidefinite for $i = 0, 1, \dots, k$.

Proof. Direct consequence of the preceding corollary using the fact that $R'R$ is positive semidefinite for any real square matrix R .

These remarks lead to an explicit characterization of stochastic controllability for systems of the form (1).

THEOREM 1. (i) *The initial state x^0 is ϵ -controllable in norm square with respect to the terminal state $x^f = 0$ in the sampling period $[0, k]$ if and only if $x^{0'}(Q(k) - \epsilon I)x^0 \leq 0$.*

(ii) *System (1) is completely ϵ -controllable with respect to the terminal state $x^f = 0$ in the sampling interval $[0, k]$ if and only if $Q(k) - \epsilon I$ is negative semidefinite.*

Proof. (i) From Lemma 1,

$$x^{0'}(Q(k) - \epsilon I)x^0 \leq 0$$

if and only if

$$x^{0'}Q(k)x^0 = \min_{u[0,k]} \mathcal{E}[\|x(k)\|^2] \leq \epsilon \|x^0\|^2 = \epsilon x^{0'}x^0,$$

where the last equality follows from the fact that we are using the Euclidean norm.

(ii) The proof in (i) holds for all x^0 if and only if $Q(k) - \epsilon I$ is negative semidefinite.

Theorem 1 indicates that a complete answer to the question of whether or not (1) is stochastically controllable in the sampling interval $[0, k]$ may be obtained through an analysis of the matrix $Q(k)$. For this reason, $Q(k)$ will be called the controllability matrix.

COROLLARY 3. *Let the stochastic sequence $\{\omega_i\}$ be statistically stationary (we may then omit reference to the time index i). Let the random scalar coefficients $\alpha_j(\omega)$, $j = 0, 1, \dots, m-1$, be independent of the sampling period (time invariant). Assume $\text{Cov}[\alpha_i(\omega), \alpha_j(\omega)] = \delta_{ij} \text{var } \alpha_i(\omega)$, where δ_{ij} is the Kronecker delta. Then the m th order scalar random difference equation*

$$x(i+m) + \alpha_{m-1}(\omega)x(i+m-1) + \dots + \alpha_0(\omega)x(i) = u(i)$$

is completely ϵ -controllable in norm square in the sense that, for any set of initial conditions $x(0) = x_0, \dots, x(m-1) = x_{m-1}$, there are a control law $u(i, x(i), \dots, x(i+m))$, $i = 0, \dots, k-1$, and a number k for which

$$\mathcal{E}[\|x(i+m+k)\|^2] \leq \epsilon(\|x_0\|^2 + \dots + \|x_{m-1}\|^2)$$

if and only if

$$\max(\text{var } \alpha_0(\omega), 1 + \text{var } \alpha_j(\omega); j = 1, \dots, m-1)$$

$$\leq \frac{\epsilon}{(1 + \text{var } \alpha_{m-1}(\omega))^{k-1}}.$$

Proof. The given equation is equivalent to

$$\begin{bmatrix} y_1(i+1) \\ \vdots \\ y_m(i+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_0(\omega) & -\alpha_1(\omega) & -\alpha_2(\omega) & \cdots & -\alpha_{m-1}(\omega) \end{bmatrix} \begin{bmatrix} y_1(i) \\ \vdots \\ y_m(i) \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 1 \end{bmatrix} u(i),$$

and, using Lemma 1, the controllability matrix is, for arbitrary k ,

$$Q(k) = [1 + \text{var } \alpha_{m-1}(\omega)]^{k-1}$$

$$\cdot \begin{bmatrix} \text{var } \alpha_0(\omega) & 0 & \cdots & 0 \\ 0 & 1 + \text{var } \alpha_1(\omega) & & \cdot \\ 0 & 0 & \cdot & \cdot \\ \vdots & \vdots & \cdot & \cdot \\ 0 & 0 & \cdots & 1 + \text{var } \alpha_{m-1}(\omega) \end{bmatrix},$$

and the corollary follows from Theorem 1 (ii).

It is well known (cf. Kalman, Ho and Narendra [11]) that the equivalent deterministic system is completely controllable. The corollary illustrates that, in the stochastic case, the question of controllability cannot be resolved solely by an analysis of the algebraic aspects of the problem; on the contrary, the probabilistic aspects may be the dominant considerations.

In the case $x^f = 0$, the control law given in Lemma 1 is essentially one for which the resulting dynamic process is a "moment reducing" process; that is, a control law which drives the second moment of the state vector at sampling period k to a minimum. It is of interest to ask for a characterization of processes of the form (1) for which the second moment of the state vector at time k can be driven to zero.

THEOREM 2. *A necessary and sufficient condition for $E[\|x(k)\|^2] = 0$ for all initial conditions $x(0)$ is that there exists a sequence of constant $m \times n$ matrices K_i , $i = 0, \cdots, k-1$, such that*

$$\prod_{i=0}^{k-1} [A_i + B_i K_i]$$

is the zero random matrix, where the matrix multiplication is carried out by pre-multiplying

$$\prod_{i=0}^j [A_i + B_i K_i] \quad \text{by} \quad [A_{j+1} + B_{j+1} K_{j+1}].$$

Proof. $\mathcal{E}\{\|x(k)\|^2\} = 0$ if and only if $x(k)$ is the zero random vector. Iterating (1) shows that

$$x(k) = A_{k-1}x(k-1) + B_{k-1}u(k-1, x(k-1)).$$

Recall that the control law which minimizes $\mathcal{E}\{\|x(k)\|^2\}$ is a linear feedback control law, i.e., $u(i, x(i)) = K_i x(i)$, $i = 0, 1, \dots, k-1$. Hence, we may restrict our attention to such control laws. Then

$$x(k) = [A_{k-1} + B_{k-1}K_{k-1}]x(k-1).$$

But $x(k-1)$ is a random variable depending on $x(k-2)$ through the same type of relation so that, by completing the recursion, the condition follows.

Note that, in the above proof, explicit use has been made of the fact that each state after the first is a stochastic variable and so also are the control vectors $u(i, x(i))$, $i = 0, 1, \dots, k-1$, since they depend on the current state of the system. The preceding development has simply written out these facts explicitly.

Note that, in order for the condition of the theorem to be satisfied, the matrix product

$$\prod_{i=0}^{k-1} [A_i + B_i K_i]$$

must contain singular matrices as subproducts. Note further that the theorem requires that this matrix product be identically zero in the random variables ω_i .

The preceding theorem, even though its criterion involves the solution of a complicated nonlinear algebraic equation, provides some interesting insight into the dynamical nature of the stochastic process under consideration. The intuitive content of the theorem is that, under certain circumstances, the stochastic elements of the process can be made to "cancel out" each other by use of the "moment reducing transformations" that have been considered thus far. It will be of interest to consider an example of Theorem 2.

Example 1.

$$\begin{bmatrix} x_1(i+1) \\ x_2(i+1) \end{bmatrix} = \begin{bmatrix} 0 & \omega_i \\ \omega_i & 0 \end{bmatrix} \begin{bmatrix} x_1(i) \\ x_2(i) \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_i \end{bmatrix} u(i), \quad x(0) = x^0,$$

where $\{\omega_i\}_{i=0}^2$ is a sequence of independent random variables with distribution functions F_i , $i = 0, 1, 2$. We ask if the second moment of the state vector can be reduced to zero in two sampling periods. Letting $K_i = [k_1(i), k_2(i)]$, Theorem 2 requires that we determine whether or not there exists a solution to

$$\begin{aligned} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} &= \prod_{i=0}^1 \left\{ \begin{bmatrix} 0 & \omega_i \\ \omega_i & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \omega_i \end{bmatrix} [k_1(i) \quad k_2(i)] \right\} \\ &= \begin{bmatrix} 0 & \omega_1 \\ [1 + k_1(1)]\omega_1 & k_2(1)\omega_1 \end{bmatrix} \begin{bmatrix} 0 & \omega_0 \\ [1 + k_1(0)]\omega_0 & k_2(0)\omega_0 \end{bmatrix} \\ &= \begin{bmatrix} [1 + k_1(0)]\omega_0\omega_1 & k_2(0)\omega_0\omega_1 \\ [1 + k_1(0)]k_2(1)\omega_0\omega_1 & \{[1 + k_1(1)] + k_2(0)k_2(1)\}\omega_0\omega_1 \end{bmatrix}. \end{aligned}$$

Let $k_1(0) = k_1(1) = -1$, $k_2(0) = k_2(1) = 0$, and the equation is solved. Thus, it is possible to reduce the second moment of the state vector to zero. In fact, it is easy to show that the required controls and controllability matrices are

$$\begin{aligned} u(1) &= -x_1(1), \\ Q(1) &= \begin{bmatrix} 0 & 0 \\ 0 & \mathcal{E}[\omega_1^2] \end{bmatrix}, \\ u(0) &= -x_1(0), \\ Q(2) &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \end{aligned}$$

and, since $Q(2) = 0$, $\mathcal{E}[\|x(2)\|^2] = 0$.

It is of interest to note that a similar system which has fewer random elements, namely the system

$$\begin{bmatrix} x_1(i+1) \\ x_2(i+1) \end{bmatrix} = \begin{bmatrix} 0 & \omega_i \\ \omega_i & 0 \end{bmatrix} \begin{bmatrix} x_1(i) \\ x_2(i) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(i),$$

does not have the property that the second moment of the state vector can be reduced to zero in any finite number of sampling periods. This example shows that the property characterized in the preceding theorem depends on both the algebraic and probabilistic structure of the problem.

4. Applications and implications of stochastic controllability. In this section, the application of the concept of stochastic controllability to particular stochastic optimal control problems is examined. In addition, the

abstract implications which the concept of ϵ -controllability in norm square holds for stochastic systems of the form (1) will be discussed.

The obvious application of controllability to stochastic optimal control is demonstrated in the following problem.

TIME OPTIMAL CONTROL PROBLEM. Given system (1) and a "stopping rule," $\mathcal{E}[\|x(k)\|^2] \leq \rho$ ($\rho = \text{const.}$) which implicitly defines the terminal sampling period k , find a control law $u_{[0, k^*)}$ such that $\mathcal{E}[\|\phi(k^*; x^0, 0, u_{[0, k^*)})\|^2] \geq \rho$ and k^* is the minimal k for which the stopping rule is satisfied.

From the definition of ϵ -controllability in norm square in §3, it is apparent that a necessary and sufficient condition for the existence of the control law $u_{[0, k^*)}$ is that the system (1) be ϵ -controllable in norm square in some sampling interval $[0, k]$ for $\epsilon = \rho/\|x^0\|^2$. If the system is ϵ -c.n.s., then it is a simple matter to determine the minimal $k = k^*$ for which the process is ϵ -c.n.s. Moreover, given k^* , the time optimal control sequence, $u_{[0, k^*)}$, is determined according to the theory given in Lemma 1. This proves the following theorem.

THEOREM 3. *A necessary and sufficient condition for the existence of an optimal control law $u_{[0, k^*)}$ for the time optimal control problem is that system (1) be ϵ -c.n.s. with respect to $x^f = 0$ in some sampling interval $[0, k]$. Moreover, given k^* , the time optimal control law is given in Lemma 1.*

It is important to note that the stochastic time optimal control problem considered here is a simplified version of the usual formulation. The usual formulation imposes certain constraints on admissible control vectors $u(i, x(i))$, $i = 0, 1, \dots, k^* - 1$. The point of the present discussion is to emphasize the relation between the qualitative concept of stochastic controllability and the quantitative concept of stochastic optimal control.

A more interesting relationship between stochastic controllability and stochastic optimal control is found in the following problem.

INFINITE HORIZON REGULATOR PROBLEM. Given system (1) (in which the sequence $\{\omega_i\}$ of random variables is now assumed to be statistically stationary) and an objective function $J = \sum_{i=0}^{\infty} \mathcal{E}[x'(i)Sx(i)]$, where S is a symmetric, positive definite matrix, find a control law $u_{[0, \infty)}$ which minimizes J .

Kalman [9] has shown that a necessary and sufficient condition for the existence of an optimal solution to the regulator problem is that there exists a control sequence $u_{[0, \infty)}$ such that $J = \sum_{i=0}^{\infty} \mathcal{E}[x'(i)Sx(i)]$ converges. Kalman's result leads to the next theorem.

THEOREM 4. *A necessary condition for the existence of an optimal solution to the infinite horizon regulator problem is that, for any $\epsilon > 0$, system (1) be ϵ -c.n.s. with respect to $x^f = 0$ for some sampling interval $[0, k]$, $k < \infty$.*

Proof.

$$\lim_{k \rightarrow \infty} \sum_{i=0}^k \mathcal{E}[x'(i)Sx(i)]$$

exists only if

$$\lim_{i \rightarrow \infty} \mathcal{E}[x'(i)Sx(i)] = \lim_{i \rightarrow \infty} \mathcal{E}[\|x(i)\|_S^2] = 0.$$

We suppose that $S \geq I$ (where $S \geq I$ if $S - I$ is positive semidefinite); then $\mathcal{E}[x'(i)Sx(i)] \geq \mathcal{E}[x'(i)x(i)]$ for all $x(i)$, and so

$$\lim_{i \rightarrow \infty} \mathcal{E}[\|x(i)\|_S^2] \geq \lim_{i \rightarrow \infty} \mathcal{E}[\|x(i)\|^2] \geq 0$$

implies

$$\lim_{i \rightarrow \infty} \mathcal{E}[\|x(i)\|^2] = 0,$$

which shows that system (1) is necessarily ϵ -c.n.s. if a solution to the infinite horizon regulator problem exists for the case $S \geq I$.

Since $S > 0$, it is always possible to find $\alpha > 0$ such that $\alpha S \geq I$, so the same argument shows that

$$\lim_{i \rightarrow \infty} \mathcal{E}[\|x(i)\|^2] = 0$$

for general $S > 0$.

This theorem is particularly useful from a practical point of view. Kalman [9] gives a number of necessary and sufficient conditions for the existence of a solution to the regulator problem, but these all depend upon finding a fixed point of a recursive matrix equation; if the dimension of the state vector x is large, this may become an impracticable computation. For this reason Theorem 4 is interesting since stochastic controllability may be established with much greater ease than a fixed point of a matrix recursion. These remarks will take on much greater significance in the next section when the focus of attention turns to a special class of systems for which the controllability question can be answered in terms of the underlying random process without recourse to the controllability matrix.

The remainder of this section will be concerned with the abstract implications that the concept of stochastic controllability holds for system (1). The first result shows how the idea of stochastic controllability induces a direct sum decomposition of initial state space, $X^0 = E^n$, into completely controllable and completely uncontrollable subspaces.

THEOREM 5. *Let x^i , $i = 1, \dots, n$, be a complete orthonormal set of eigenvectors of the matrix $Q(k) - \epsilon I$. Let $S \subset E^n$ be the subspace spanned by the given orthonormal eigenvectors which correspond to nonpositive eigenvalues of $Q(k) - \epsilon I$. Let $S^\perp \subset E^n$ be the subspace spanned by the given orthonormal eigenvectors corresponding to positive eigenvalues of $Q(k) - \epsilon I$. Then*

- (i) $X^0 = S \oplus S^\perp$;
- (ii) *all initial states $x \in S$ are ϵ -c.n.s. with respect to $x^f = 0$ in the sampling interval $[0, k]$;*

(iii) *no initial state* $x \in S^+$ *is* ϵ -c.n.s. *with respect to* $x^f = 0$ *in the sampling interval* $[0, k]$.

Proof. (i) Follows from the definition of S and S^+ .

(ii) Let $x^i, i = 1, \dots, s$, be the orthonormal eigenvectors of $Q(k) - \epsilon I$ which are a basis for S and let $x^i, i = s + 1, \dots, n$, be the orthonormal eigenvectors of $Q(k) - \epsilon I$ which are a basis for S^+ . Note that the existence of the orthogonal eigenvectors $x^i, i = 1, \dots, n$, is guaranteed by the fact that $Q(k) - \epsilon I$ is a real symmetric matrix. Let Γ be the matrix whose i th column is $x^i, i = 1, \dots, n$. Then Γ diagonalizes $Q(k) - \epsilon I$:

$$\Gamma \Lambda \Gamma' = Q(k) - \epsilon I,$$

where $\Lambda = ((\delta_{ij}\lambda_i))$, δ_{ij} is the Kronecker delta and λ_i is the i th eigenvalue of $Q(k) - \epsilon I$. Note that, from the ordering of the $x^i, i = 1, \dots, n$, in the matrix Γ , the first s λ_i are nonpositive and the last $(n - s)$ λ_i are positive. Now consider $z = \sum_{i=1}^s c_i x^i \in S$, where $c_i, i = 1, \dots, s$, are real constants. We have

$$\begin{aligned} z'(Q(k) - \epsilon I)z &= z'\Gamma\Lambda\Gamma'z = \left(\sum_{i=1}^s c_i x^i\right)' \Gamma\Lambda\Gamma' \left(\sum_{i=1}^s c_i x^i\right) \\ &= \sum_{i=1}^s c_i^2 \lambda_i \\ &\leq 0, \end{aligned}$$

since $\lambda_i \leq 0, i = 1, \dots, s$. Hence,

$$z'(Q(k) - \epsilon I)z \leq 0,$$

and, by Theorem 1, statement (ii) holds.

(iii) Same argument as in (ii) with the obvious modifications.

It is important to notice that the set $C(0, k; \epsilon)$ of states which are ϵ -controllable with respect to the origin in the sampling interval $[0, k]$ is not a linear subspace (as it is in the deterministic case) but that $C(0, k; \epsilon) \supset S$. This is because there are states which are ϵ -c.n.s. (i.e., states in $C(0, k; \epsilon)$) which have components both in S and S^+ (i.e., which are not in S or S^+). Another way of saying this is that $x'(Q(k) - \epsilon I)x$ is nonpositive for all initial states, x , in $C(0, k; \epsilon)$ but such x may be direct sums of vectors in S and S^+ .

The last result in this section will depend on the assumptions that

(i) $A(i, \omega_i)$ is a deterministic quantity, that is, $A(i, \omega_i) = A(i)$, $i = 0, 1, \dots, k - 1$;

(ii) $\det A(i) \neq 0, i = 0, 1, \dots, k - 1$.

Assumption (ii) implies that $\det \Phi(k, i) \neq 0, i = 0, 1, \dots, k - 1$, and hence that system (1) is reversible.

LEMMA 3 (cf. [11] for deterministic version). *Under assumptions (i) and*

(ii),

$$\phi(0; C(k, h; \epsilon), k, 0) \subset C(0, h; \epsilon) \quad \text{for } h \geq k.$$

Proof. x^0 is in $C(0, h; \epsilon)$ if $\phi(k; x^0, 0, 0)$ is in $C(k, h; \epsilon)$ for $h \geq k$.

5. A special class of stochastic systems. It was noted in §3 that a complete answer to whether or not system (1) is stochastically controllable with respect to $x^f = 0$ in the sampling interval $[0, k]$ can be obtained if the controllability matrix $Q(k)$ is known. Unfortunately, if the dimension, n , of (1) is large and k is large, the calculation of $Q(k)$ is time-consuming. In this section we will characterize a subclass of system (1) for which the question of stochastic controllability can be resolved entirely in terms of the algebraic and probabilistic properties of $A(i, \omega_i)$ and $B(i, \omega_i)$. This will lead to simple necessary conditions and sufficient conditions for controllability.

The main assumptions necessary for the development of the following theory are:

(i) the sequence $\{\omega_i\}_{i=0}^\infty$ is a sequence of independent, identically distributed random variables, and thus we may omit explicit reference to the time period i ;

(ii) $A(i, \omega)$ and $B(i, \omega)$ are time-invariant random matrices, and thus $A(i, \omega) = A(\omega)$ and $B(i, \omega) = B(\omega)$;

(iii) the random variable ω has positive probability at only a finite number of points $\omega^j, j = 1, \dots, M$, and has mass function $\Pr[\omega = \omega^j] = p_j, j = 1, \dots, M$;

(iv) the matrices $A(\omega)$ and $B(\omega)$ have elements in an algebraically closed field \mathfrak{F} ;

(v) $m = n$ so that $B(\omega)$ is an $n \times n$ random matrix.

Under assumption (ii), the matrix $\Psi(i), i = 0, 1, \dots$, introduced in Corollary 1 has M realizations corresponding to the M realizations of the stationary random variable ω . Denote these realizations by $\Psi^j, j = 1, \dots, M$; similarly denote the realizations of $A(\omega)$ and $B(\omega)$ by $A^j, B^j, j = 1, \dots, M$.

It should be noted that assumption (v) is merely a formality; it is necessary for technical reasons which will be noted later. It is acceptable to allow the last $n - m$ columns of $B(\omega)$ to be the zero n -vectors; this effectively reduces $B(\omega)$ to dimension $n \times m, m \leq n$. Note, however, that if $B(\omega)$ is an $n \times m, m < n$, matrix, it is essential that it be extended to an $n \times n$ matrix in the manner mentioned, i.e., let the last $n - m$ columns be zero vectors. Any other extension of $B(\omega)$ to an $n \times n$ matrix having $n - m$ zero columns does not guarantee that the dynamical properties of the system remain the same as when $B(\omega)$ was an $n \times m, m < n$, matrix. Analogous remarks apply if $m > n$.

We will need the following definition.

DEFINITION 4. Let \mathfrak{F} denote an algebraically closed field. Then the $n \times n$ matrices $S^j, j = 1, \dots, M$, with elements in \mathfrak{F} have Property P if and only if the characteristic roots $\lambda_i, i = 1, \dots, n$, of every matrix polynomial $f(S^1, \dots, S^M)$ with coefficients in \mathfrak{F} are all of the form $\lambda_i = f(\lambda_i^1, \dots, \lambda_i^M)$, where λ_i^j is the i th characteristic root of $S^j, j = 1, \dots, M$, for some ordering $i = 1, \dots, n$.

The presence of property P is useful since it allows one to compute the eigenvalues of a (matrix) polynomial by evaluating the corresponding (scalar) polynomial with each matrix argument replaced by a corresponding eigenvalue. This property will be used heavily in the sequel.

We shall also need the following lemma.

LEMMA 4. Let $S^j, j = 1, \dots, M$, have Property P and let $f(S^1, \dots, S^M)$ be a symmetric matrix polynomial with coefficients in the field \mathfrak{F} . Then the eigenvalues $\lambda_i^\dagger, i = 1, \dots, n$, of the pseudoinverse $f(S^1, \dots, S^M)^\dagger$ of every such matrix polynomial have the form

$$\lambda_i^\dagger = \begin{cases} f(\lambda_i^1, \dots, \lambda_i^M)^{-1} & \text{if } f(\lambda_i^1, \dots, \lambda_i^M) \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

where λ_i^j is the i th eigenvalue of $S^j, j = 1, \dots, M$, for the same ordering $i = 1, \dots, n$, of the eigenvalues of each S^j as that required for Property P.

Proof. The hypothesis that $f(S^1, \dots, S^M)$ is symmetric implies the existence of an orthogonal matrix Γ such that

$$f(S^1, \dots, S^M) = \Gamma' \Theta \Gamma,$$

where

$$\Theta_{ij} = \theta_i \delta_{ij},$$

where δ_{ij} is the Kronecker delta and $\theta_i, i = 1, \dots, n$, are the eigenvalues of $f(S^1, \dots, S^M)$. Then, by the discussion of the pseudoinverse in §3,

$$f(S^1, \dots, S^M)^\dagger = \Gamma' \Theta^\dagger \Gamma,$$

where Θ^\dagger is diagonal and has the eigenvalues of $f(S^1, \dots, S^M)^\dagger$ on the diagonal. We have, from §3,

$$\Theta_{ii}^\dagger = \begin{cases} \theta_i^{-1} & \text{if } \theta_i \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

But the eigenvalues $\theta_i, i = 1, \dots, n$, of $f(S^1, \dots, S^M)$ are, by Property P,

$$\theta_i = f(\lambda_i^1, \dots, \lambda_i^M),$$

where λ_i^j is the i th eigenvalue of $S^j, j = 1, \dots, M$.

It is of interest to note that the preceding lemma does not require that the symmetric matrix polynomial $f(S^1, \dots, S^M)$ have an inverse. This will be of significant interest to us in our attempt to maintain the foregoing level of generality.

Using the preceding material, it is possible to develop an efficient method for calculating the eigenvalues of the matrix $Q(k)$ described in §3. The result is the following lemma.

LEMMA 5. *Suppose the set of matrices consisting of the realizations of the system matrices and their transposes, $A^j, A^{j'}, B^j$ and $B^{j'}, j = 1, \dots, M$, has Property P. Then*

$$\lambda_i[Q(k)] = (\lambda_i[Q(1)])^k.$$

Proof. Let λ_i^j and $\bar{\lambda}_i^j$, $i = 1, \dots, n$, denote the eigenvalues of A^j and $A^{j'}$, $j = 1, \dots, M$. Let μ_i^j and $\bar{\mu}_i^j$, $j = 1, \dots, M$, denote the eigenvalues of B^j and $B^{j'}$, $j = 1, \dots, M$. We assume that the ordering of the eigenvalues is fixed and corresponds to the ordering for which $A^j, A^{j'}, B^j$ and $B^{j'}$, $j = 1, \dots, M$, have Property P. Also, we denote the i th eigenvalue of a matrix R by $\lambda_i[R]$. Separate notation for the eigenvalues of A^j and $A^{j'}$ and B^j and $B^{j'}$ is used to emphasize that a very special ordering $i = 1, \dots, n$ of the eigenvalues is being used. In this ordering, it may not be true that $\lambda_i[A^j] = \lambda_i[A^{j'}]$ or that $\lambda_i[B^j] = \lambda_i[B^{j'}]$.

The proof is by induction. The first step in the induction is easy and is identical to the general case; hence, we give only the inductive proof for general k . If $\lambda_i[Q(1)] = 0$, the result holds trivially. Suppose $\lambda_i[Q(1)] \neq 0$ and that the result holds for $k - 1$ and try to extend the induction:

$$\begin{aligned} \lambda_i[Q(k)] &= \lambda_i\{\varepsilon[A'Q(k-1)A] \\ &\quad - \varepsilon[A'Q(k-1)B]\varepsilon[B'Q(k-1)B]^\dagger\varepsilon[B'Q(k-1)A]\} \\ &= \lambda_i\left[\sum_{j=1}^M p_j A^{j'}Q(k-1)A^j - \left(\sum_{j=1}^M p_j A^{j'}Q(k-1)B^j\right) \right. \\ &\quad \cdot \left.\left(\sum_{j=1}^M p_j B^{j'}Q(k-1)B^j\right)^\dagger \left(\sum_{j=1}^M p_j B^{j'}Q(k-1)A^j\right)\right]. \end{aligned}$$

Noting (i) $Q(k-1)$ is a (symmetric) matrix polynomial in $A^j, A^{j'}, B^j$ and $B^{j'}$, $j = 1, \dots, M$, and therefore under the assumptions made, $Q(k-1)$ has Property P, (ii) Lemma 4, (iii) the induction hypothesis, and (iv) the fact that, since the eigenvalues of $A^j, A^{j'}, B^j$ and $B^{j'}$, $j = 1, \dots, M$, are elements of the field \mathfrak{F} , they commute with each other, we may write

$$\begin{aligned} \lambda_i[Q(k)] &= (\lambda_i[Q(1)])^{k-1} \left[\sum_{j=1}^M p_j \bar{\lambda}_i^j \lambda_i^j \right. \\ &\quad \left. - \left(\sum_{j=1}^M p_j \bar{\mu}_i^j \lambda_i^j \right) \left(\sum_{j=1}^M p_j \bar{\mu}_i^j \mu_i^j \right)^{-1} \left(\sum_{j=1}^M p_j \bar{\lambda}_i^j \mu_i^j \right) \right] \\ &= (\lambda_i[Q(1)])^{k-1} \lambda_i[Q(1)] \\ &= (\lambda_i[Q(1)])^k. \end{aligned}$$

Lemma 5 leads to the next theorem.

THEOREM 6. *Let assumptions (i)–(v) of this section hold. Let the realizations $A^j, A^{j'}, B^j, B^{j'}, j = 1, \dots, M$, of the random matrices $A(\omega), A'(\omega), B(\omega)$ have Property P. Then a necessary and sufficient condition for system (1) to be completely ϵ -controllable in norm square with respect to the terminal state $x^f = 0$ in the sampling interval $[0, k]$ is that*

$$(5) \quad \max_{1 \leq i \leq n} \lambda_i \{ \mathcal{E}[\Psi'(\omega)\Psi(\omega)] \} \leq \epsilon^{1/k},$$

where $\Psi(\omega)$ is the matrix introduced in Corollary 1.

Proof. We recall from Theorem 1 that system (1) is completely ϵ -c.n.s. if and only if $x^{0'}Q(k)x^0 \leq \epsilon x^{0'}x^0$ for all initial states x^0 . This condition is equivalent to

$$(6) \quad \max_{x^0} \frac{x^{0'}Q(k)x^0}{x^{0'}x^0} \geq \epsilon.$$

We now show that (6) holds if and only if condition (5) holds. We know (see [1, p. 110]) that

$$\max_{x^0} \frac{x^{0'}Q(k)x^0}{x^{0'}x^0} = \max_{1 \leq i \leq n} \lambda_i [Q(k)].$$

Under the hypotheses of the theorem, Lemma 5 shows that

$$\lambda_i [Q(k)] = (\lambda_i [Q(1)])^k = (\lambda_i \{ [\Psi'(\omega)\Psi(\omega)] \})^k.$$

Thus, (6) holds if and only if condition (5) holds.

It should be noted that assumption (v) came into play in Theorem 6 in an essential manner. That is, $B(\omega)$ must be an $n \times n$ matrix in order for it to have well defined eigenvalues; hence assumption (v) is necessary.

At this point, the presence of Property P has already allowed us to obtain some useful results. However, the presence of Property P is obviously difficult to verify and so we seek alternative (equivalent) forms of Property P.

Now¹ let \mathcal{Q} be the algebra of polynomials in the matrices $A^j, A^{j'}, B^j, B^{j'}, j = 1, \dots, M$. The elements of \mathcal{Q} then form a representation of the algebra \mathcal{Q} . Hence, it is known (see [18, Chap. X]) that all elements of \mathcal{Q} may, by a similarity transformation, be simultaneously reduced to the form

¹ The following two paragraphs provide the necessary background for the remaining theorems in this section; the discussion is abstracted from [14]. A good general introduction to the ideas in this section may be found in [18, Chap. X].

$$(7) \quad \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1s} \\ 0 & B_{22} & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \\ \cdot & & & \cdot & \cdot \\ 0 & \cdots & & B_{ss} \end{bmatrix},$$

where the set $\mathfrak{A}^{(k)}$ of square matrices B_{kk} , $k = 1, 2, \dots, s$, is an irreducible representation of \mathfrak{A} . These irreducible components $\mathfrak{A}^{(k)}$ of \mathfrak{A} are uniquely determined to within a similarity transformation and thus their orders are completely determined. The fact that $\mathfrak{A}^{(k)}$ is an irreducible representation of \mathfrak{A} means essentially that $\mathfrak{A}^{(k)}$ is a simple algebra, i.e., one with no proper invariant subalgebra. We shall assume without loss of generality that all elements of \mathfrak{A} are in the form (7).

Let \mathfrak{R} denote the minimum invariant subalgebra of \mathfrak{A} containing all the elements

$$S^i S^j - S^j S^i, \quad i, j = 1, 2, \dots, M,$$

where $S^i = A^i, A^{i'}, B^i$ or $B^{i'}$, $i = 1, \dots, M$. If C_i , $i = 1, 2, \dots, s$, is a basis of \mathfrak{R} , the general element of \mathfrak{R} may be written in the form

$$\mathfrak{R}_x = \sum_{i=1}^s x_i C_i,$$

where the x_i are indeterminate elements of \mathfrak{F} . The characteristic polynomial $\det(\mathfrak{R}_x - \lambda)$ of \mathfrak{R} is called the characteristic polynomial of \mathfrak{R} .

Using Theorem 6, we can now develop results which characterize stochastic controllability in terms of the algebraic and probabilistic properties of $A(\omega)$ and $B(\omega)$. We shall use the following hypothesis.

HYPOTHESIS 1. We shall say that Hypothesis 1 holds if and only if one of the following (equivalent) hypotheses holds.

- (i) Exactly n of the irreducible components $\mathfrak{A}^{(k)}$ of \mathfrak{A} are of order 1.
- (ii) The characteristic polynomial of \mathfrak{R} is divisible by λ^n .
- (iii) The quotient ring $\mathfrak{A}/\mathfrak{R}$ where \mathfrak{R} is the radical of \mathfrak{A} (i.e., maximum nilpotent invariant subalgebra of \mathfrak{A}) is commutative.
- (iv) The field \mathfrak{F} has characteristic 0 and the Lie algebra defined by the matrices $A^j, A^{j'}, B^j, B^{j'}, j = 1, \dots, M$, is solvable.
- (v) There exists a nonsingular matrix T such that $T^{-1}S^i T$ simultaneously triangularizes $S^i = A^i, A^{i'}, B^i$ or $B^{i'}$, $i = 1, \dots, M$.

The equivalence of hypotheses (i)–(v) is proved in McCoy [14].

Remark. Comparing McCoy [14] and Hoffman and Taussky [7], we find that Hypothesis 1 implies that A^i and B^i are normal matrices (i.e.,

P is normal if and only if $P'P = PP'$. Then (cf. Gantmacher [6, p. 272]) each $A^{i'}$ can be represented as a polynomial in A^i and similarly for $B^{i'}$. Hence, it is evident that it is sufficient to restrict Hypothesis 1 to the set $\tilde{S}^i = A^i, B^i, i = 1, \dots, M$.

We let $\alpha_i (\beta_i)$ denote the rows of $A(\omega) (B(\omega))$ and $\alpha^i (\beta^i)$ denote the columns of $A(\omega) (B(\omega))$, and then we can state the next theorem.

THEOREM 7. *If Hypothesis 1 and assumptions (i)–(v) of this section hold and if the columns $\beta^i, i = 1, \dots, n$, of $B(\omega)$ are pairwise orthogonal random vectors, then a sufficient condition for system (1) to be completely ϵ -controllable in norm square with respect to the terminal state $x^f = 0$ in the sampling interval $[0, k]$ is*

$$(8) \quad \max_{1 \leq i \leq n} \sum_{j=1}^n \left| \mathcal{E}[\alpha^i \cdot \alpha^j] - \sum_{t \in \mathfrak{J}} \frac{\mathcal{E}[\alpha^i \cdot \beta^t] \mathcal{E}[\beta^t \cdot \alpha^j]}{\mathcal{E}[\beta^t \cdot \beta^t]} \right| \leq \epsilon^{1/k},$$

where $\mathfrak{J} = \{t: \mathcal{E}[\beta^t \cdot \beta^t] \neq 0\}$.

Proof. McCoy [14] proves that Property P holds if and only if Hypothesis 1 holds. Then, by Theorem 6, a sufficient condition for system (1) to be ϵ -c.n.s. is

$$\max_{1 \leq i \leq n} \lambda_i \{ \mathcal{E}[\Psi'(\omega) \Psi(\omega)] \} \leq \epsilon^{1/k}.$$

By a theorem of Frobenius (see [4, p. 75]),

$$\max_{1 \leq i \leq n} \lambda_i \{ \mathcal{E}[\Psi'(\omega) \Psi(\omega)] \} \leq \max_{1 \leq i \leq n} \sum_{j=1}^n | \mathcal{E}[\Psi'(\omega) \Psi(\omega)]_{ij} |.$$

We now evaluate $\mathcal{E}[\Psi'(\omega) \Psi(\omega)]_{ij}$. From Lemma 1,

$$\begin{aligned} \mathcal{E}[\Psi'(\omega) \Psi(\omega)]_{ij} &= Q_{ij}(1) \\ &= \mathcal{E}[A'(\omega) A(\omega)]_{ij} - \{ \mathcal{E}[A'(\omega) B(\omega)] \mathcal{E}[B'(\omega) B(\omega)]^\dagger \mathcal{E}[B'(\omega) A(\omega)] \}_{ij}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{E}[B'(\omega) A(\omega)]_{ij} &= \mathcal{E}[\beta^i \cdot \alpha^j], \\ \mathcal{E}[A'(\omega) B(\omega)]_{ij} &= \mathcal{E}[\alpha^i \cdot \beta^j], \end{aligned}$$

so that

$$\begin{aligned} & \{ \mathcal{E}[A'(\omega) B(\omega)] \mathcal{E}[B'(\omega) B(\omega)]^\dagger \mathcal{E}[B'(\omega) A(\omega)] \}_{ij} \\ &= \sum_{t=1}^m \mathcal{E}[A'(\omega) B(\omega)]_{it} \sum_{s=1}^m \mathcal{E}[B'(\omega) B(\omega)]_{ts}^\dagger \mathcal{E}[B'(\omega) A(\omega)]_{sj} \\ &= \sum_{t \in \mathfrak{J}} \frac{\mathcal{E}[\alpha^i \cdot \beta^t] \mathcal{E}[\beta^t \cdot \alpha^j]}{\mathcal{E}[\beta^t \cdot \beta^t]}, \end{aligned}$$

and so

$$\varepsilon[\Psi'(\omega)\Psi(\omega)]_{ij} = [\alpha^i \cdot \alpha^j] - \sum_{t \in \mathfrak{J}} \frac{\varepsilon[\alpha^i \cdot \beta^t] \varepsilon[\beta^t \cdot \alpha^j]}{\varepsilon[\beta^t \cdot \beta^t]},$$

from which the result follows.

Remark. It is apparent that the assumption that the columns of $B(\omega)$ are pairwise orthogonal is not restrictive; and analogous result holds for the general case.

If we rewrite the left-hand side of condition (8) as

$$\max_{1 \leq i \leq n} \sum_{j=1}^m \left| \varepsilon[\alpha^i \cdot \alpha^j] \left(1 - \sum_{t \in \mathfrak{J}} \frac{\varepsilon[\alpha^i \cdot \beta^t] \varepsilon[\beta^t \cdot \alpha^j]}{\varepsilon[\alpha^i \cdot \alpha^i] \varepsilon[\beta^t \cdot \beta^t]} \right) \right|$$

and interpret the quantity

$$\sum_{t \in \mathfrak{J}} \frac{\varepsilon[\alpha^i \cdot \beta^t] \varepsilon[\beta^t \cdot \alpha^j]}{\varepsilon[\alpha^i \cdot \alpha^i] \varepsilon[\beta^t \cdot \beta^t]}$$

as a “correlation coefficient” between $A(\omega)$ and $B(\omega)$, condition (8) has the interpretation that if the “correlation coefficient” is too small, $B(\omega)$ does not “synchronize” with $A(\omega)$ and the control is ineffective.

THEOREM 8. *Under the same hypotheses as in Theorem 7, a necessary condition for system (1) to be completely ϵ -c.n.s. with respect to the terminal state $x^f = 0$ in the sampling interval $[0, k]$ is that*

$$(9) \quad \min_{1 \leq i \leq n} \left| \left(\left| \varepsilon[\alpha^i \cdot \alpha^i] - \sum_{t \in \mathfrak{J}} \frac{(\varepsilon[\alpha^i \cdot \beta^t])^2}{\varepsilon[\beta^t \cdot \beta^t]} \right| - \sum_{\substack{j=1 \\ j \neq i}}^n \left| \varepsilon[\alpha^i \cdot \alpha^j] - \sum_{t \in \mathfrak{J}} \frac{\varepsilon[\alpha^i \cdot \beta^t] \varepsilon[\beta^t \cdot \alpha^j]}{\varepsilon[\beta^t \cdot \beta^t]} \right| \right) \right| \leq \epsilon^{1/k},$$

where \mathfrak{J} is defined in Theorem 7.

Proof. By a theorem of Frobenius (see [4, p. 75]),

$$\begin{aligned} \min_{1 \leq i \leq n} \lambda_i \{ \varepsilon[\Psi'(\omega)\Psi(\omega)] \} \\ \geq \min_{1 \leq i \leq n} \left| \left(\left| \varepsilon[\Psi'(\omega)\Psi(\omega)]_{ii} \right| - \sum_{\substack{j=1 \\ j \neq i}}^n \left| \varepsilon[\Psi'(\omega)\Psi(\omega)]_{ij} \right| \right) \right|. \end{aligned}$$

If system (1) is completely ϵ -c.n.s., condition (5) of Theorem 6 holds and

$$\begin{aligned} \epsilon^{1/k} &\geq \max_{1 \leq i \leq n} \lambda_i \{ \varepsilon[\Psi'(\omega)\Psi(\omega)] \} \\ &\geq \min_{1 \leq i \leq n} \lambda_i \{ \varepsilon[\Psi'(\omega)\Psi(\omega)] \} \\ &\geq \min_{1 \leq i \leq n} \left| \left(\left| \varepsilon[\Psi'(\omega)\Psi(\omega)]_{ii} \right| - \sum_{\substack{j=1 \\ j \neq i}}^n \left| \varepsilon[\Psi'(\omega)\Psi(\omega)]_{ij} \right| \right) \right|. \end{aligned}$$

The evaluation of $\mathcal{E}[\Psi'(\omega)\Psi(\omega)]_{ij}$ given in the proof of Theorem 7 yields the condition (9).

Theorems 7 and 8 give, respectively, sufficient and necessary conditions for complete ϵ -controllability in norm square. These conditions are stated in terms of the linear algebraic and probabilistic properties of the system (1) and so, from a computational point of view, are much easier to examine than is the controllability matrix.

6. Controllability in adaptive systems. In this section, the problem of controllability in discrete adaptive dynamical decision processes is considered. The difference between this class of problems and those previously considered will be in the form of the information pattern. In the stochastic control processes considered in previous sections, a complete description of the state of the process was composed of:

(1) a point, $x(i)$, in state space;

(2) an information pattern, \mathcal{I} , which was composed of knowledge of the distribution function(s) of the underlying random variable(s).

The description of the state in the adaptive system is composed of (1) but the information pattern in (2) is modified. More particularly, the adaptive aspect of the problem is introduced through the assumption that we do not completely know the distribution function $F(\omega)$ of the independent, identically distributed random variables ω_i , $i = 0, 1, \dots$. That is, we know the functional form of $F(\omega)$ but we do not know some of its parameters. Bayesian analysis will be used to provide a solution to this problem.

Within this general framework, the problem will be the same as in previous sections: given a time-invariant system of the form (1), where $\{\omega_i\}$ is a sequence of independent, identically distributed random variables with distribution function $F(\omega)$ having unknown parameters, determine a necessary and sufficient condition for ϵ -c.n.s. It will be seen that the adaptive control process possesses a more dynamic nature (in a certain sense) than the stochastic problem previously considered. This will be due to the fact that the distribution of the random variable ω is not completely known and the a priori estimate of it must be modified as more information is gained about the process. This is done as the process progresses in time.

Contrast this with the stochastic control problem in which complete information on the stochastic variable is available at time 0. In this case, the stochastic controllability of the system at time k can be characterized by the controllability matrix $Q(k)$ since all of the moments used in constructing $Q(k)$ are known (since the distribution of ω is known). Thus, in the stochastic control process, the question of controllability (or lack of it) can be resolved at time 0 from the knowledge of the distribution of ω and the dynamics of the system.

On the other hand, the information pattern of the adaptive process

evolves along with the process. We now indicate how the information pattern, \mathcal{I} , in the adaptive process evolves along with the process. We suppose that we possess, at sampling period i , an a priori distribution function $G(i, \alpha)$ of the vector of parameters α of the distribution function $F(\omega)$ of the random variable ω . This, in turn, gives an a priori distribution function $F(\omega | G(i, \alpha))$ of the random variable ω at time i . We also suppose that we possess a procedure for modifying this a priori estimate on the basis of the actual state resulting from the decision $u(i)$, namely $x(i+1)$, and the information we already possess, namely $x(i)$, $u(i)$ and $G(i, \alpha)$. Thus, as a result of a decision $u(i)$,

$$x(i) \rightarrow x(i+1),$$

$$G(i, \alpha) \rightarrow G(i+1, \alpha | x(i+1), x(i), u(i), G(i, \alpha)).$$

This notation indicates that the new a priori estimates depend on the new state, the previous state and the decision made. The form of this transformation will be dictated, as will be shown by example, by Bayes' rule. The system is then in state $x(i+1)$ and possesses a new information pattern $G(i+1, \alpha)$. This leads to a new a priori distribution function for ω , namely, $F(\omega | G(i+1, \alpha))$. The fact that only $F(\omega | G(i, \alpha))$ is available at sampling period i means that the controllability of the process cannot be determined completely at time 0. For this reason, we introduce the concept of adaptive controllability at time i_0 .

DEFINITION 5 (cf. Definition 1). An initial state x^0 is ϵ -controllable in norm square in the adaptive sense with respect to the terminal state $x^f = 0$ in the sampling interval $[i_0, i_0 + k]$ if and only if there exists a control law $u_{[i_0, i_0+k]} = \{u(i, x(i)), i = i_0, i_0 + 1, \dots, i_0 + k - 1\}$ such that

$$E[\|\phi(i_0 + k; x^0, i_0, u_{[i_0, i_0+k]})\|^2] \leq \epsilon \|x^0\|^2,$$

where the expectation is taken with respect to the joint distribution of the independent, identically distributed random variables ω_i , $i = i_0, \dots, i_0 + k - 1$, conditioned on the information pattern, \mathcal{I} , available at sampling period i_0 .

With this background, there is no difficulty in writing down the functional equation which describes the evolution of the process. Let

$f_{i_0+k-1}(x)$ = minimum value of $E[\|x(i_0 + k)\|^2]$ over the $i_0 + k - 1$ remaining sampling periods of the process, starting in state x with information pattern $G(i, \alpha)$ at sampling period i , $i_0 \leq i \leq i_0 + k$, subject to (1) and using an optimal policy for the remaining sampling periods,

and we obtain the same functional equation as in §2, where the expectation is now with respect to the a priori distribution of ω , namely $F(\omega | G(i, \alpha))$. The same techniques used in previous sections are valid, and Lemma 2

holds for the adaptive case where the expectations in Lemma 2 are now taken with respect to the prior distribution of ω at sampling period i . Since the special case of system (1) under consideration is time-invariant, Lemma 2 also holds when the terminal sampling period is $i_0 + k$ and the initial sampling period is i_0 .

An example will clarify the preceding remarks and will show how Bayes' rule is used to update $F(\omega | G(i, \alpha))$.

Example 2. Consider the system

$$\begin{bmatrix} x_1(i+1) \\ x_2(i+1) \end{bmatrix} = \begin{bmatrix} 0 & \omega \\ \omega & 0 \end{bmatrix} \begin{bmatrix} x_1(i) \\ x_2(i) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(i), \quad x(0) = x,$$

where ω is a random variable which can take on two values, $+\frac{1}{2}$ and $-\frac{1}{2}$, and

$$f(\omega) = p^{\omega+1/2}(1-p)^{1/2-\omega}, \quad \omega = -\frac{1}{2}, \frac{1}{2}.$$

Thus, knowledge of p completely specifies the distribution of ω . If we possessed this knowledge, we would have the stochastic control problem discussed previously. In the present (adaptive) case, we assume that we do not know p but do have an a priori distribution function $G(0, p)$ of p at time 0.

We first compute the return function and control policy in terms of the moments of ω . In view of Definition 5, the expectation is to be conditioned on our knowledge of the random process ω . If we knew the distribution of ω (i.e., if we knew p), the expectations would be with respect to the distribution of ω and we would have the case studied in previous sections. On the other hand, when we do not know the distribution of ω , the expectation is conditioned on the information pattern, $\mathcal{I}(i)$, of the process which is available at the sampling period, i , at which the expectation is evaluated. This yields, in general, for an arbitrary number of stages (this is valid since we have assumed that the sequence of random variables is identically distributed)

$$u(i) = -[\mathcal{E}[\omega | \mathcal{I}(i)] \quad 0]x(i),$$

$$Q(i+1) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} Q(i) \begin{bmatrix} 0 & \mathcal{E}[\omega^2 | \mathcal{I}(i)] \\ \text{var}[\omega | \mathcal{I}(i)] & 0 \end{bmatrix}.$$

Thus we see that, in order to discuss the stochastic controllability of this system, we only need knowledge of $\mathcal{E}[\omega^2 | \mathcal{I}(i)]$ and $\text{var}[\omega | \mathcal{I}(i)] = \mathcal{E}[\omega^2 - \mathcal{E}^2[\omega | \mathcal{I}(i)] | \mathcal{I}(i)]$.

We now ask the question: "Given the information pattern of the system

at some time i_0 , can ϵ -c.n.s. be achieved after k transitions, i.e., at time $i_0 + k$?" Note that this is a proper question to ask since the question of controllability of the system is not dependent upon the state of the system; it is only a property of the dynamic and stochastic properties of the system.

Now, using the a priori distribution of p , $G(0, p)$, we show how to answer this question. First we note that, if we have no other information on the history of the stochastic process ω at time 0, the information pattern is comprised solely of the prior distribution $G(0, p)$ at time 0. Hence, at 0, if we wish to find

$$\mathcal{E}[u(\omega) \mid \mathcal{I}(0)],$$

where $u(\omega) = \omega^2$ or $\omega^2 - \mathcal{E}[\omega \mid \mathcal{I}(0)]$, we find

$$\mathcal{E}[u(\omega) \mid G(0, p)].$$

To evaluate this expectation, we must find $f(\omega \mid G(0, p))$; this is obtained as

$$f(\omega \mid G(0, p)) = \int f(\omega \mid p(G)) dG(0, p),$$

and hence

$$\begin{aligned} \mathcal{E}[u(\omega) \mid G(0, p)] &= \int u(\omega) f(\omega \mid G(p)) d\omega \\ &= \int u(\omega) \int f(\omega \mid p) dG(0, p) d\omega. \end{aligned}$$

If, at sampling period 0, we observe that $\omega = \frac{1}{2}$, the posterior (Bayes) estimate of the probability for p will be

$$dH(p \mid 1, 0, G) = \frac{p dG(0, p)}{\int_0^1 p dG(0, p)},$$

and, after observing $\omega = -\frac{1}{2}$, the posterior distribution will be

$$dH(p \mid 0, 1, G) = \frac{(1-p) dG(0, p)}{\int_0^1 (1-p) dG(0, p)}.$$

In summary, if the a priori choice of a distribution for p is $G(0, p)$ and if, at time $i_0 = m + n$, we have observed $\omega = \frac{1}{2}$ m times and $\omega = -\frac{1}{2}$ n times, the posterior (Bayes) estimate of the distribution of p is

$$dH(p \mid m, n, G) = \frac{p^m (1-p)^n dG(0, p)}{\int_0^1 p^m (1-p)^n dG(0, p)}.$$

Note that $dH(p | m, n, G)$ summarizes the information pattern at time $i_0 = m + n$ since it incorporates all relevant information which has been gathered about the process in $m + n$ sampling periods together with the information available prior to time 0 (since all prior information is presumably summarized in $G(0, p)$).

Hence, at time $i_0 = m + n$, if we wish to find

$$\mathcal{E}[u(\omega) | \mathcal{G}(i_0)],$$

we find

$$\mathcal{E}[u(\omega) | m, n, G] = \mathcal{E}[u(\omega) | H(p | m, n, G)].$$

To evaluate this expectation, we must find $f(\omega | m, n, G)$; this is found as

$$f(\omega | m, n, G) = \int f(\omega | p(m, n, G)) dH(p | m, n, G),$$

and hence

$$\mathcal{E}[u(\omega) | m, n, G] = \int u(\omega) \int f(\omega | p(m, n, G)) dH(p | m, n, G).$$

To complete the example, assume that $dG(0, p) = g(p)$ is a Beta density function

$$g(p) = \frac{p^{a-1}(1-p)^{b-1}}{B(a, b)},$$

where $B(a, b)$ is the Beta function,

$$dH(p | m, n, G) = h(p | m, n, dG) = \frac{p^{m+a-1}(1-p)^{n+b-1}}{B(m+a, n+b)},$$

that is, $h(p | m, n, g)$ is a Beta density function. The parameters a and b play the roles of the a priori numbers of positive and negative realizations of ω that are observed. If the sum $a + b$ is small, not much weight is given to the initial estimate; if it is large, many periods are required before the estimation of p can be significantly changed.

Then we have (cf. Raiffa and Schlaifer [17, p. 237])

$$f(\omega | m, n, g) = \frac{1}{(m+n+a+b)(m+a-1)!(n+b-1)!} \\ \cdot \left(m+a+\omega-\frac{1}{2}\right)! \left(n+b+\omega-\frac{1}{2}\right)!,$$

$$\mathcal{E}[\omega^2 | m, n, g] = \frac{1}{4},$$

$$\text{var } [\omega | m, n, g] = \frac{4(m+a)(n+b)}{[(m+a) + (n+b)]^2}.$$

Now it is possible to illustrate the concept of adaptive controllability. Suppose we are at time i_0 and wish to determine the controllability of the system at time $i_0 + 4$ using the information pattern available at i_0 . Since we have assumed stationarity, we have that the return function evaluated four stages in the future is given by

$$Q(4) = \begin{bmatrix} (\varepsilon[\omega^2 | m, n, g])^2 (\text{var } [\omega | m, n, g])^2 & 0 \\ 0 & (\varepsilon[\omega^2 | m, n, g])^2 (\text{var } [\omega | m, n, g])^2 \end{bmatrix}.$$

We assume a value for ϵ and assume that $i_0 = m + n$ where we have observed $\omega = +\frac{1}{2}m$ times and $\omega = -\frac{1}{2}n$ times since the start of the process so that we have $i_0 = m + n$ observations in the information pattern. Then

$$Q(4) - \epsilon I = \begin{bmatrix} \frac{(m+a)(n+b)}{4[(m+a)+(n+b)]^2} - \epsilon & 0 \\ 0 & \frac{(m+a)(n+b)}{4[(m+a)+(n+b)]^2} - \epsilon \end{bmatrix},$$

and the ϵ -controllable subspaces are determined as in the stochastic control process. It is also apparent that all of the analysis employed in the discussion of the stochastic control process can be used in the analysis of the present adaptive control process.

7. Acknowledgment. Most of the ideas of this paper are taken directly from the writer's dissertation [5]. Many thanks are due to Professor R. E. Kalman for having originally suggested the problem.

REFERENCES

- [1] R. E. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
- [2] ———, *Adaptive Control Processes*, Princeton University Press, Princeton, 1960.
- [3] R. E. BELLMAN AND R. KALABA, *On adaptive control processes*, IRE Trans. Automatic Control, AC-4 (1959), pp. 1-10.
- [4] E. BODEWIG, *Matrix Calculus*, North-Holland, Amsterdam, 1959.
- [5] M. M. CONNORS, *Controllability of discrete, linear, random dynamical systems*, Doctoral dissertation, Stanford University, Stanford, 1966.
- [6] F. R. GANTMACHER, *Matrix Theory*, Chelsea, New York, 1960.
- [7] A. J. HOFFMAN AND O. TAUSKY, *A characterization on normal matrices*, J. Res. Nat. Bur. Standards Sect. B., 52 (1954), pp. 17-19.
- [8] R. E. KALMAN, *On the general theory of control systems*, Proceedings of the First International Congress on Automatic Control, Moscow, 1960, Butterworth's Scientific Publications, vol. 1, London, 1961, pp. 481-492.
- [9] ———, *Control of randomly varying linear dynamical systems*, Symposia on Applied Mathematics, vol. 13, American Mathematical Society, New York, 1962, pp. 287-298.
- [10] ———, *Mathematical description of linear dynamical systems*, this Journal, 1 (1963), pp. 152-192.

- [11] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1(1963), pp. 189–213.
- [12] R. E. KALMAN AND T. S. ENGLAR, *A User's Manual for ASP C*, Contract NAS 2-1107, Research Institute of Advanced Study, Baltimore, 1966.
- [13] N. N. KRASOVSKII, *On a problem in tracking*, J. Appl. Math. Mech., 26 (1962), pp. 314–335.
- [14] N. H. MCCOY, *On the characteristic roots of matrix polynomials*, Bull. Amer. Math. Soc., 42 (1936), pp. 592–600.
- [15] R. PENROSE, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.
- [16] ———, *On best approximate solutions of linear matrix equations*, Ibid., 52 (1965), pp. 17–19.
- [17] H. RAIFFA AND R. SCHLAIFER, *Applied statistical decision theory*, Division of Research, Harvard Business School, Boston, 1961.
- [18] J. H. M. WEDDERBURN, *Lectures on Matrices*, Dover, New York, 1964.

CONSTRUCTION OF THE REGION OF CONTROLLABILITY FOR SYSTEMS WITH BOUNDED-IMPULSE CONTROLS*

A. M. FORMAL'SKII†

Abstract. For a system described by linear differential equations the problem is solved of constructing a region in the state space, from each point of which the origin can be reached by means of a bounded-impulse control.

Consider the system described by a linear vector differential equation with constant, real coefficients,

$$(1) \quad \frac{dx}{dt} = Ax + Bu.$$

Here $x = \|x_j\|$ is an $n \times 1$ matrix, $A = \|a_{ij}\|$ is an $n \times n$ matrix, $B = \|b_j\|$ is an $n \times 1$ matrix, u is a scalar time function.

As admissible controls we shall take measurable functions $u(\tau)$ satisfying the inequality

$$(2) \quad \int_0^T |u(\tau)| d\tau \leq C,$$

where $C = \text{const.}$, while T is the time such that $u(\tau) \equiv 0$ when $\tau > T$. We denote the set of admissible controls by Ω .

In every real physical system, $u(t)$ is a generalized force. From a physical point of view inequality (2) implies the boundedness of the impulse of the control.

Let the roots of the characteristic equation

$$(3) \quad \det \|A - \lambda E\| = 0$$

(E is the identity matrix) be $\lambda_k = \epsilon_k + i\omega_k$ of multiplicity p_k , where the λ_k have positive real parts for $k = 1, \dots, r_1$, zero real parts for $k = r_1 + 1, \dots, r_2$, and negative real parts for $k = r_2 + 1, \dots, r_3$. In addition to the roots indicated, to every complex root there corresponds a complex conjugate root, so that $\sum_{k=1}^{r_3} \zeta_k p_k = n$, where $\zeta_k = 1$ if λ_k is a real root and $\zeta_k = 2$ if λ_k is a complex root.

The general solution of system (1) has the form

$$(4) \quad x(t) = N(t)x_0 + \int_0^t N(t)N^{-1}(\tau)Bu(\tau) d\tau,$$

* Originally published in Vestnik Moskovskogo Universiteta, No. 5 (1966), pp. 75-84. Submitted on May 18, 1965. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Department of Applied Mechanics, Moscow State University, Moscow.

where $N(t)$ is the fundamental matrix of the solutions of the homogeneous system¹ ($N(0) = E$), while x_0 is the initial state vector.

We pose the problem of determining in the state space X the set Q of initial states x_0 from which the system can be brought to the origin by means of an admissible control.

If the relation $x(T) = 0$ holds for some value of T , then after multiplying (4) on the left by $N^{-1}(T)$ at $t = T$ we obtain

$$(5) \quad -x_0 = \int_0^T N^{-1}(\tau) Bu(\tau) d\tau.$$

The well-known equality $N(t)N^{-1}(\tau) = N(t - \tau)$ holds for a system with constant coefficients. Setting $t = 0$ in this equality we obtain $N(0)N^{-1}(\tau) = N(-\tau)$; but $N(0) = E$ and, consequently, $N^{-1}(\tau) = N(-\tau)$. Replacing $N^{-1}(\tau)$ by $N(-\tau)$ in (5) we obtain

$$(6) \quad -x_0 = \int_0^T N(-\tau) Bu(\tau) d\tau.$$

The stated problem can be reformulated thus: to determine the region Q of vectors x_0 for which (6) is ensured by means of an admissible control.

We introduce the notation

$$(7) \quad v(T) = \int_0^T N(-\tau) Bu(\tau) d\tau.$$

In the state space X we consider a set Q_T of points, the endpoints of the vectors $v(T)$ which may be obtained at the instant T under all possible admissible controls $u(\tau)$ defined on the interval $[0, T]$. The set Q_T has the following properties:

- (i) the set Q_T is convex;
- (ii) if $v \in Q_{T_1}$ and $T_2 > T_1$, then $v \in Q_{T_2}$;
- (iii) the set Q_T is symmetric with respect to the origin.

We prove Property (i). Let $v_1(T) \in Q_T$ and $v_2(T) \in Q_T$ under the controls $u_1(\tau)$ and $u_2(\tau)$, respectively. Consider the equation $u^*(\tau) = \beta u_1(\tau) + (1 - \beta)u_2(\tau)$, where $0 \leq \beta \leq 1$. The function $u^*(\tau)$ is measurable and satisfies inequality (2); consequently, the control $u^*(\tau)$ is admissible. From (7) we have $v^*(T) = \beta v_1(T) + (1 - \beta)v_2(T)$. An appropriate choice of the magnitude of β can be made for every point of the segment connecting the points $v_1(T)$ and $v_2(T)$. Property (i) is proved. The convexity of the set Q_T can also be proved by the use of the theory of moments [1]. We now prove Property (ii). Let $v(T_1) \in Q_{T_1}$ under an admissible control $u(\tau)$, and let

¹ *Translator's note:* In modern terminology this matrix is often called the transition matrix, and this latter term is used throughout the rest of the article.

$T_2 > T_1$. We define a control $u^*(\tau)$ in the following way:

$$u^*(\tau) = \begin{cases} u(\tau) & \text{for } \tau \leq T_1, \\ 0 & \text{for } T_1 < \tau \leq T_2. \end{cases}$$

It is obvious that the control $u^*(\tau)$ is admissible and that $v(T_1) = v^*(T_2) \in Q_{T_2}$. Property (iii) follows from the fact that if in (7) we replace the control $u(\tau)$ by the control $-u(\tau)$, also admissible, then instead of $v(T)$ we get $-v(T)$.

From the definition of set Q_T and from Property (iii) it follows that system (1) can be brought to the origin in time T from only those points which belong to Q_T . The set Q , which was to be determined, is by virtue of Property (ii) the set of points in state space which includes Q_T as $T \rightarrow \infty$. Let us show that the set Q is convex. Let $v_1 \in Q$ and $v_2 \in Q$; this means that there exist T_1 and T_2 such that $v_1 \in Q_{T_1}$ and $v_2 \in Q_{T_2}$. To be specific, let $T_2 > T_1$; then from Property (ii) it follows that $v_1 \in Q_{T_2}$, and in such case, by virtue of Property (i), all the points of the segment connecting v_1 and v_2 belong to the set Q_{T_2} , and hence, also to the set Q .

We take a unit vector η and we draw the hyperplanes of support of the set Q_T , orthogonal to the vector η (Fig. 1). By virtue of Property (i) there will be two such planes; in view of Property (iii), they will be symmetric with respect to the origin. Let us determine the distance $d_\eta(\tau)$ from the origin to these planes. We multiply (7) on the left by the row-matrix η :

$$\eta v(T) = \int_0^T \eta N(-\tau) B u(\tau) d\tau.$$

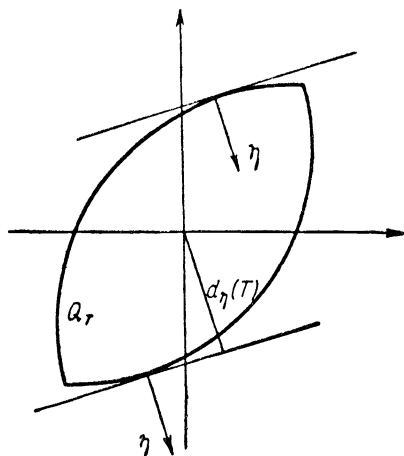


FIG. 1

We then obtain an expression (rigorously proved in [2]) which is geometrically obvious:

$$(8) \quad d_{\eta}(T) = \max_{v(T) \in Q_T} (\eta v(T)) = \max_{u(\tau) \in \Omega} \int_0^T \eta N(-\tau) B u(\tau) d\tau.$$

Let us show that the control which satisfies the equality $\int_0^T |u(\tau)| d\tau = C$ ensures the maximum of the integral $\int_0^T \eta N(-\tau) B u(\tau) d\tau$. Indeed, let us assume that the maximizing function $u(\tau)$ satisfies the inequality $\int_0^T |u(\tau)| d\tau < C$, but then by choosing $\alpha > 1$ such that $\int_0^T |\alpha u(\tau)| d\tau = C$, we obtain $\int_0^T \eta N(-\tau) B \alpha u(\tau) d\tau > \int_0^T \eta N(-\tau) B u(\tau) d\tau$, since $\int_0^T \eta N(-\tau) \cdot B u(\tau) d\tau > 0$.

Thus, to determine $d_{\eta}(T)$ and the corresponding function $u(\tau)$ we must solve the following problem: maximize the functional

$$(9) \quad I(u) = \int_0^T \eta N(-\tau) B u(\tau) d\tau$$

under the condition

$$(10) \quad \int_0^T |u(\tau)| d\tau = C.$$

We solve this problem in the following way.

The function $|\eta N(-\tau) B|$ is continuous in the interval $[0, T]$; consequently, it attains its maximum in this interval. Let the maximum be reached for $\tau = \tau_1$:

$$(11) \quad M_T(\eta) = \max_{\tau \in [0, T]} |\eta N(-\tau) B| = |\eta N(-\tau_1) B|.$$

For any function $u(\tau)$ satisfying (10), we have

$$(12) \quad \begin{aligned} I(u) &= \int_0^T \eta N(-\tau) B u(\tau) d\tau \\ &\leq \int_0^T |\eta N(-\tau) B| |u(\tau)| d\tau \leq C M_T(\eta). \end{aligned}$$

Consider the sequence of step functions

$$(13) \quad u_m(\tau) = \begin{cases} \frac{Cm}{2\Delta} \operatorname{sgn} [\eta N(-\tau) B] & \text{when } \tau \in \left[\tau_1 - \frac{\Delta}{m}, \tau_1 + \frac{\Delta}{m} \right], \\ 0 & \text{when } \tau \notin \left[\tau_1 - \frac{\Delta}{m}, \tau_1 + \frac{\Delta}{m} \right], \end{cases}$$

where Δ is a sufficiently small quantity. Each of the functions $u_m(\tau)$ satisfies (10), and the sequence (13) defines the delta-function $C \operatorname{sgn} [\eta N(-\tau_1)B] \cdot \delta(\tau - \tau_1)$ which is nonzero at the point $\tau = \tau_1$. Using the theorem on the mean, from (9) we obtain

$$\begin{aligned} I(u_m) &= \int_0^T \eta N(-\tau) B u_m(\tau) d\tau \\ &= \int_{\tau_1 - \Delta/m}^{\tau_1 + \Delta/m} |\eta N(-\tau) B| \frac{Cm}{2\Delta} d\tau \\ &= |\eta N(-\tau_1 + \xi) B| C, \end{aligned}$$

where $|\xi| \leq \Delta/m$. Hence, taking (11) into account, we obtain

$$(14) \quad \lim_{m \rightarrow \infty} I(u_m) = CM_T(\eta).$$

Thus, from (8), (12) and (14), follows

$$(15) \quad d_\eta(T) = \max_{u(\tau) \in \Omega} I(u) = CM_T(\eta),$$

and, moreover, the maximum of the functional $I(u)$ is achieved when $u(\tau) = C \operatorname{sgn} [\eta N(-\tau_1)B] \delta(\tau - \tau_1)$. (The fact that the control which drives system (1) to the origin in minimum time is a delta-function was shown in [3].)

As is well known (for example, see [4]), every element of the matrix $N(-\tau)$ has the form

$$N_{ij}(-\tau) = \sum_{k=1}^{r_3} \sum_{l=1}^{p_k} e^{-\epsilon_k \tau} \tau^{p_k-l} (\alpha_{\lambda_k l}^{ij} \cos \omega_k \tau + \beta_{\lambda_k l}^{ij} \sin \omega_k \tau), \quad i, j = 1, \dots, n,$$

where $\alpha_{\lambda_k l}^{ij}$ and $\beta_{\lambda_k l}^{ij}$ are constants. If λ_k is a real root, $\omega_k = 0$ and $\beta_{\lambda_k l}^{ij} = 0$. The expression for $\eta N(-\tau)B$ has the form

$$(16) \quad \begin{aligned} \eta N(-\tau)B &= \sum_{k=1}^{r_3} \sum_{l=1}^{p_k} e^{-\epsilon_k \tau} \tau^{p_k-l} \\ &\cdot \left[\left(\sum_{i=1}^n \eta_i \alpha_{\lambda_k l}^i \right) \cos \omega_k \tau + \left(\sum_{i=1}^n \eta_i \beta_{\lambda_k l}^i \right) \sin \omega_k \tau \right], \end{aligned}$$

where

$$\begin{aligned} \alpha_{\lambda_k l}^i &= \sum_{j=1}^n b_j \alpha_{\lambda_k l}^{ij}, \\ \beta_{\lambda_k l}^i &= \sum_{j=1}^n b_j \beta_{\lambda_k l}^{ij}, \end{aligned} \quad l = 1, \dots, p_k, \quad k = 1, \dots, r_3, \quad i = 1, \dots, n.$$

Let us consider the system of linear algebraic equations

$$(17) \quad \sum_{i=1}^n \eta_i \alpha_{\lambda_k l}^i = 0, \quad l = 1, \dots, p_k - 1 \quad \text{for} \quad k = r_1 + 1, \dots, r_2,$$

$$\sum_{i=1}^n \eta_i \beta_{\lambda_k l}^i = 0, \quad l = 1, \dots, p_k \quad \text{for} \quad k = r_2 + 1, \dots, r_3.$$

This system contains $\sum_{k=r_1+1}^{r_2} \zeta_k(p_k - 1) + \sum_{k=r_2+1}^{r_3} \zeta_k p_k$ equations. To (17) let us adjoin the normalizing condition

$$(18) \quad \eta_1^2 + \dots + \eta_n^2 = 1.$$

Any solution η^0 of (17), (18) when substituted into (16) annihilates all the terms containing $e^{-\epsilon_{r_2+1}\tau}, \dots, e^{-\epsilon_{r_3}\tau}$, where $\epsilon_{r_2+1} < 0, \dots, \epsilon_{r_3} < 0$, and terms not containing exponentials but containing τ^{p_k-l} , where $p_k - l \geq 1$. Thus, the function $|\eta^0 N(-\tau)B|$ remains bounded as $\tau \rightarrow \infty$; consequently, the function $M_T(\eta^0) = \max_{\tau \in [0, T]} |\eta^0 N(-\tau)B|$ tends to a finite limit as $T \rightarrow \infty$. We denote $\lim_{T \rightarrow \infty} M_T(\eta^0)$ by $M(\eta^0)$ and $CM(\eta^0)$ by d_{η^0} . It is obvious that $M(-\eta^0) = M(\eta^0)$ and $d_{-\eta^0} = d_{\eta^0}$. In the case $\eta \neq \eta^0$, the quantity $M_T(\eta) \rightarrow \infty$ as $T \rightarrow \infty$. Let η^0 be a vector such that $d_{\eta^0} \neq 0$; then from what has been said above it follows that the set Q is contained between the planes

$$(19) \quad \eta^0 x = d_{\eta^0},$$

$$(20) \quad -\eta^0 x = d_{\eta^0}.$$

For certain vectors η^0 , for example, for those which the expression $\eta^0 N(-\tau)B$ contains only exponential terms with negative powers, the function $M_T(\eta^0)$ is independent of T for values of T larger than some T' ; consequently, $M(\eta^0) = M_{T'}(\eta^0)$. For such η^0 , as T increases, the set Q_T reaches the planes (19) and (20) when $T = T'$ and does not "extend" any further in the direction of η^0 for a subsequent increase in T . Thus, for such η^0 there exist on the planes (19) and (20) points belonging to the set Q , and the points of set Q satisfy the inequality

$$(21) \quad |\eta^0 x| \leq d_{\eta^0}.$$

For those η^0 for which the function $M_T(\eta^0)$ increases steadily as T increases, the set Q_T reaches the planes (19) and (20) only when $T \rightarrow \infty$, and the coordinates of the points of set Q satisfy the inequality

$$(22) \quad |\eta^0 x| < d_{\eta^0}.$$

The quantity d_{η^0} can be equal to zero only for those vectors η^0 which satisfy the system of n algebraic equations

$$(23) \quad \sum_{i=1}^n \eta_i \alpha_{\lambda_k l}^i = 0, \quad \sum_{i=1}^n \eta_i \beta_{\lambda_k l}^i = 0, \quad l = 1, \dots, p_k, \quad k = 1, \dots, r_3.$$

Let ρ be the rank of system (23); then the fundamental system of solutions of (23) consists of $n - \rho$ vectors. Let us enumerate each of these vectors and denote them by $\eta^{(1)}, \dots, \eta^{(n-\rho)}$. Then, obviously, the set Q lies in the planes

$$(24) \quad \eta^\sigma x = 0, \quad \sigma = 1, \dots, n - \rho,$$

i.e., the dimension of the set Q equals ρ .

Let the vectors

$$(25) \quad B, AB, \dots, A^{n-1}B$$

be linearly independent. In this case, as is well known, $\rho = n$. Indeed, the row-matrix $\eta N(-\tau)$ is a solution of the adjoint system of equations

$$\frac{dz}{d\tau} = -zA,$$

with the initial conditions $z(0) = \eta$. It was shown in [5] that if $z(\tau)$ is a nontrivial solution of the adjoint system, then, under condition (25), $z(\tau)B$ vanishes a finite number of times in any interval $[0, \tau]$. From this it follows that system (23) cannot have a nonzero vector η as a solution, and hence, $\rho = n$. For $\rho = n$ the equations in (17) are independent since system (17) is a part of system (23).

The set of points in the space X which satisfy conditions (21), (22), (24) is denoted by Q' . Then Q is contained in the set Q' . Let us now show that every point of set Q' belongs to Q . From the definition of the set Q' it follows that for any plane passing through an arbitrary point x belonging to Q' , we can find a T such that the set Q_T , and hence also Q , contains the points of this plane. The set Q is convex and, therefore, if the point x did not belong to Q , then through it we could pass a plane lying entirely outside Q , which is impossible. Thus, Q coincides with Q' .

The equations and the inequalities obtained allow us to ascertain the structure of the set Q . The set Q is a cylindrical set; the dimension of the base of this cylinder equals the dimension of the fundamental system of solutions of (17). In case condition (25) is satisfied, this dimension equals $n - \sum_{k=r_1+1}^{r_2} \zeta_k(p_k - 1) - \sum_{k=r_2+1}^{r_3} \zeta_k p_k$. The part of the boundary of set Q satisfying conditions (21) belongs to Q ; the other part, satisfying (22), does not belong to the set Q . The results obtained allow us to construct the set Q in any specific case.

Let us consider two particular cases, assuming that condition (25) is satisfied.

1. All the roots of (3) have negative real parts. In this case the system of equations (17) coincides with system (23) which contains n equations and which, by virtue of (25), has only the trivial solution. Consequently,

for all $\eta \neq 0$, the quantity $d_\eta(T) \rightarrow \infty$ as $T \rightarrow \infty$, and hence, the region of controllability occupies the whole state space.

2. All the roots of (3) with the exception of λ_1 have negative real parts. The root λ_1 is either a zero root of multiplicity p_1 or a real positive root of multiplicity $p_1 = 1$. In this case system (17) consists of $n - 1$ linearly-independent equations. Equations (17) and (18) have only two solutions, differing from each other in sign, η^0 and $-\eta^0$. From (11), (15) and (16) we have

$$d_{\eta^0} = C \lim_{T \rightarrow \infty} \max_{\tau \in [0, T]} \left| \sum_{i=1}^n \eta_i^0 \alpha_{\lambda_{i1}}^i e^{-\epsilon_1 \tau} \right| = C \left| \sum_{i=1}^n \eta_i^0 \alpha_{\lambda_{i1}}^i \right|.$$

The region Q is the set of state space points bounded by two planes orthogonal to the vector η^0 and lying at a distance d_{η^0} from the origin (these planes belong to Q).

Example 1. Consider the second order system of equations

$$(26) \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = -x_1 + 2\epsilon x_2 + u,$$

where $0 < \epsilon < 1$. The matrices A and B are

$$A = \begin{Bmatrix} 0 & 1 \\ -1 & 2\epsilon \end{Bmatrix}, \quad B = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}.$$

System (26) has the two complex-conjugate roots $\epsilon \pm i\sqrt{1 - \epsilon^2}$ with positive real part. The transition matrix has the form

$$N(\tau) = \begin{Bmatrix} e^{\epsilon\tau} \left(\cos \sqrt{1 - \epsilon^2}\tau - \frac{\epsilon}{\sqrt{1 - \epsilon^2}} \sin \sqrt{1 - \epsilon^2}\tau \right) \\ -e^{\epsilon\tau} \frac{1}{\sqrt{1 - \epsilon^2}} \sin \sqrt{1 - \epsilon^2}\tau \\ e^{\epsilon\tau} \frac{1}{\sqrt{1 - \epsilon^2}} \sin \sqrt{1 - \epsilon^2}\tau \\ e^{\epsilon\tau} \left(\cos \sqrt{1 - \epsilon^2}\tau + \frac{\epsilon}{\sqrt{1 - \epsilon^2}} \sin \sqrt{1 - \epsilon^2}\tau \right) \end{Bmatrix}.$$

System (26) has no roots with nonpositive real parts and, therefore, equations of type (17) do not exist for it. It is easy to verify that condition (25) is satisfied for the system being considered. Consequently, Q is a two-dimensional set of points whose coordinates satisfy inequality (21) for all unit vectors η .

The function $\eta N(-\tau)B$ has the form

$$\eta N(-\tau)B = e^{-\epsilon\tau} \left(\eta_2 \cos \sqrt{1 - \epsilon^2}\tau - \frac{\eta_1 + \epsilon\eta_2}{\sqrt{1 - \epsilon^2}} \sin \sqrt{1 - \epsilon^2}\tau \right).$$

The maximum of the function $|\eta N(-\tau)B|$ in $0 \leq \tau \leq \infty$ is reached either at $\tau = 0$ or at the smallest τ for which the derivative of the function $\eta N(-\tau)B$ vanishes. Introducing the notations $\eta_1 = \cos \phi$, $\eta_2 = \sin \phi$, after simple computations we obtain (we drop the upper index in the vector η^0)

$$d_\eta = \begin{cases} Ce^{-\epsilon\tau_1} \sqrt{1 + \epsilon \sin 2\phi} & \text{when } e^{-\epsilon\tau_1} \sqrt{1 + \epsilon \sin 2\phi} \geq |\sin \phi|, \\ C|\sin \phi| & \text{when } e^{-\epsilon\tau_1} \sqrt{1 + \epsilon \sin 2\phi} \leq |\sin \phi|, \end{cases}$$

where

$$\tau_1 = \begin{cases} \frac{1}{\sqrt{1 - \epsilon^2}} \tan^{-1} \frac{(\cos \phi + 2\epsilon \sin \phi) \sqrt{1 - \epsilon^2}}{\epsilon \cos \phi + (2\epsilon^2 - 1) \sin \phi} & \text{when } \frac{\cos \phi + 2\epsilon \sin \phi}{\epsilon \cos \phi + (2\epsilon^2 - 1) \sin \phi} \geq 0, \\ \frac{1}{\sqrt{1 - \epsilon^2}} \left[\tan^{-1} \frac{(\cos \phi + 2\epsilon \sin \phi) \sqrt{1 - \epsilon^2}}{\epsilon \cos \phi + (2\epsilon^2 - 1) \sin \phi} + \pi \right] & \text{when } \frac{\cos \phi + 2\epsilon \sin \phi}{\epsilon \cos \phi + (2\epsilon^2 - 1) \sin \phi} < 0, \\ \frac{1}{\sqrt{1 - \epsilon^2}} \frac{\pi}{2} & \text{when } \epsilon \cos \phi + (2\epsilon^2 - 1) \sin \phi = 0. \end{cases}$$

By constructing sufficiently many straight lines of the one-parameter family

$$x_1 \cos \phi + x_2 \sin \phi = d_\eta,$$

we can obtain the boundary L of set Q with sufficient accuracy by using inequality (21). Fig. 2 shows the results of such a construction for $\epsilon = \sqrt{2}/2$, $C = 1$.

Example 2. Consider the third order system of equations

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= a_{22}x_2 + a_{23}x_3 + b_2u, \\ \dot{x}_3 &= a_{32}x_2 + a_{33}x_3 + b_3u. \end{aligned} \quad (27)$$

Equations (27) describe, for example, the lateral motion of an aircraft (x_1 is the path, x_3 is the sideslip angle). If the aircraft is controlled by reactive jets, then u is a quantity proportional to the thrust. The condition for the boundedness of the fuel consumption in such jets is expressed by inequality (2).

The characteristic equation of the homogeneous system of equations, obtained from (27) when $u \equiv 0$, has only one zero root. Let two other roots be real such that $\lambda_1 > 0$, $\lambda_2 = 0$, $\lambda_3 < 0$.

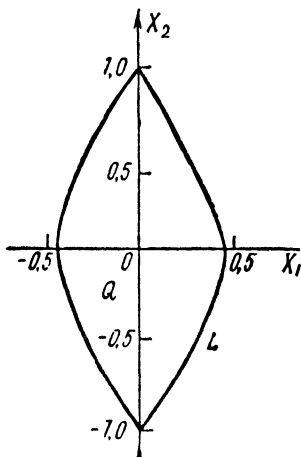


FIG. 2

The transition matrix is

$$N(\tau) = \begin{vmatrix} 1 & \alpha_{\lambda_1}^{(1,2)} e^{\lambda_1 \tau} + \alpha_{\lambda_2}^{(1,2)} + \alpha_{\lambda_3}^{(1,2)} e^{\lambda_3 \tau} & \alpha_{\lambda_1}^{(1,3)} e^{\lambda_1 \tau} + \alpha_{\lambda_2}^{(1,3)} + \alpha_{\lambda_3}^{(1,3)} e^{\lambda_3 \tau} \\ 0 & \alpha_{\lambda_1}^{(2,2)} e^{\lambda_1 \tau} + \alpha_{\lambda_3}^{(2,2)} e^{\lambda_3 \tau} & \alpha_{\lambda_1}^{(2,3)} e^{\lambda_1 \tau} + \alpha_{\lambda_3}^{(2,3)} e^{\lambda_3 \tau} \\ 0 & \alpha_{\lambda_1}^{(3,2)} e^{\lambda_1 \tau} + \alpha_{\lambda_3}^{(3,2)} e^{\lambda_3 \tau} & \alpha_{\lambda_1}^{(3,3)} e^{\lambda_1 \tau} + \alpha_{\lambda_3}^{(3,3)} e^{\lambda_3 \tau} \end{vmatrix},$$

where

$$\begin{aligned} \alpha_{\lambda_1}^{(1,2)} &= \frac{\lambda_1 - a_{33}}{\lambda_1(\lambda_1 - \lambda_3)}, & \alpha_{\lambda_1}^{(1,3)} &= -\frac{(\lambda_1 - a_{33})(\lambda_3 - a_{33})}{a_{32}\lambda_1(\lambda_1 - \lambda_3)}, \\ \alpha_{\lambda_2}^{(1,2)} &= -\frac{a_{32}}{\lambda_1\lambda_3}, & \alpha_{\lambda_2}^{(1,3)} &= \frac{(\lambda_1 - a_{33})(\lambda_3 - a_{33})}{a_{32}\lambda_1\lambda_3}, \\ \alpha_{\lambda_3}^{(1,2)} &= -\frac{\lambda_3 - a_{33}}{\lambda_3(\lambda_1 - \lambda_3)}, & \alpha_{\lambda_3}^{(1,3)} &= \frac{(\lambda_1 - a_{33})(\lambda_3 - a_{33})}{a_{32}\lambda_3(\lambda_1 - \lambda_3)}, \\ \alpha_{\lambda_1}^{(2,2)} &= \frac{\lambda_1 - a_{33}}{\lambda_1 - \lambda_3}, & \alpha_{\lambda_1}^{(3,2)} &= -\alpha_{\lambda_3}^{(3,2)} = \frac{a_{33}}{\lambda_1 - \lambda_3}, \\ \alpha_{\lambda_3}^{(2,2)} &= -\frac{\lambda_3 - a_{33}}{\lambda_1 - \lambda_3}, & \alpha_{\lambda_1}^{(3,3)} &= -\frac{\lambda_3 - a_{33}}{\lambda_1 - \lambda_3}, \\ \alpha_{\lambda_1}^{(2,3)} &= -\alpha_{\lambda_3}^{(2,3)} & \alpha_{\lambda_3}^{(3,3)} &= \frac{\lambda_1 - a_{33}}{\lambda_1 - \lambda_3}. \\ & & & = -\frac{(\lambda_1 - a_{33})(\lambda_3 - a_{33})}{a_{32}(\lambda_1 - \lambda_3)}, \end{aligned}$$

In the case being considered $p_1 = p_2 = p_3 = 1$ and, therefore, the second lower index in the coefficients of the transition matrix has been dropped.

The expression for $\eta N(-\tau)B$ has the form

$$(28) \quad \eta N(-\tau)B = e^{-\lambda_1 \tau} (\eta_1 \alpha_{\lambda_1}^{(1)} + \eta_2 \alpha_{\lambda_1}^{(2)} + \eta_3 \alpha_{\lambda_1}^{(3)}) + \eta_1 \alpha_{\lambda_2}^{(1)} + e^{-\lambda_3 \tau} (\eta_1 \alpha_{\lambda_3}^{(1)} + \eta_2 \alpha_{\lambda_3}^{(2)} + \eta_3 \alpha_{\lambda_3}^{(3)}),$$

where

$$\alpha_{\lambda_k}^{(i)} = b_2 \alpha_{\lambda_k}^{(i,2)} + b_3 \alpha_{\lambda_k}^{(i,3)}, \quad i = 1, 2, 3 \quad \text{when} \quad k = 1, 3, \\ i = 1 \quad \text{when} \quad k = 2.$$

In this example, (17) and (18) have the forms

$$(29) \quad \eta_1 \alpha_{\lambda_3}^{(1)} + \eta_2 \alpha_{\lambda_3}^{(2)} + \eta_3 \alpha_{\lambda_3}^{(3)} = 0,$$

$$(30) \quad \eta_1^2 + \eta_2^2 + \eta_3^2 = 1.$$

Equations (29) and (30) are satisfied by all possible unit vectors η^0 orthogonal to the vector $g = (\alpha_{\lambda_3}^{(1)}, \alpha_{\lambda_3}^{(2)}, \alpha_{\lambda_3}^{(3)})$. It follows from (28) that the maximum $M(\eta^0)$ of the function $|\eta^0 N(-\tau)B|$ in $0 \leq \tau \leq \infty$ is reached either at $\tau = 0$ or at $\tau = \infty$, and then we obtain

$$(31) \quad d_{\eta^0} = CM(\eta^0) = \begin{cases} C |\eta_1^0 \alpha_{\lambda_1}^{(1)} + \eta_2^0 \alpha_{\lambda_1}^{(2)} + \eta_3^0 \alpha_{\lambda_1}^{(3)} + \eta_1^0 \alpha_{\lambda_2}^{(1)}| \\ \text{when } |\eta_1^0 \alpha_{\lambda_2}^{(1)}| \leq |\eta_1^0 \alpha_{\lambda_1}^{(1)} + \eta_2^0 \alpha_{\lambda_1}^{(2)} + \eta_3^0 \alpha_{\lambda_1}^{(3)} + \eta_1^0 \alpha_{\lambda_2}^{(1)}|, \\ C |\eta_1^0 \alpha_{\lambda_2}^{(1)}| \\ \text{when } |\eta_1^0 \alpha_{\lambda_2}^{(1)}| \geq |\eta_1^0 \alpha_{\lambda_1}^{(1)} + \eta_2^0 \alpha_{\lambda_1}^{(2)} + \eta_3^0 \alpha_{\lambda_1}^{(3)} + \eta_1^0 \alpha_{\lambda_2}^{(1)}|. \end{cases}$$

Thus, the set Q is a cylinder whose axis is directed along the vector g , while the distances up to the generatrices are determined by (31).

REFERENCES

- [1] E. F. BECKENBACH AND R. BELLMAN, *Inequalities*, 2nd ed., Springer-Verlag, New York, 1965.
- [2] J. H. EATON, *An iterative solution to time-optimal control*, J. Math. Anal. Appl., 5 (1962), pp. 329-344.
- [3] N. N. KRASOVSKII, *On the theory of optimum control*, J. Appl. Math. Mech., 23 (1959), pp. 899-919.
- [4] B. V. BULGAKOV, *Oscillations*, Gostekhizdat, Moscow, 1954.
- [5] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

A PROBLEM OF OPTIMAL JOINT CONTROL*

V. B. GINDEST†

Abstract. A problem of optimal joint control of a linear system is solved for the case when the players have nonantagonistic interests, using the methods of the calculus of variations and of functional analysis.

1. Statement of the problem. We consider the problem of a nonantagonistic, noncooperative, joint control of a linear plant by two players. Each player strives to optimize his own performance criterion. We shall assume that the players cannot enter into an agreement. Each player's choice (of his control function) must be determined over the whole preassigned process interval; moreover, Player I chooses first, knowing only the goal and the capabilities of Player II, while the latter selects his control next, knowing what his opponent has already chosen.

The control system is described by the linear differential equation

$$(1) \quad \dot{x} = Ax + bu + cv, \quad x(0) = x_0,$$

where $x = x(t)$ is the n -dimensional state vector of the plant; $u = u(t)$ and $v = v(t)$ are the scalar control functions of the first and second player, respectively; the coefficient matrices A , b , c are continuous functions of time, of dimension $n \times n$, $n \times 1$, $n \times 1$, respectively. The process is considered in the time interval $[0, T]$, where $T > 0$ is the given process time.

The goal of the first player is to minimize the distance (norm) of the terminal state of the system from an assigned point r in state space, i.e., to minimize the functional

$$J_1(u, v) = \|x(T) - r\|_1,$$

where $\|y\|_1$ denotes an arbitrary norm of the element y in the n -dimensional state space X_n of the control system.

The goal of the second player is to minimize the functional [1]

$$J_2(u, v) = \|x(T)\|_2^2 + \lambda \int_0^T v^2(\tau) d\tau.$$

Here $\|x(T)\|_2$ is the Euclidean norm ($\|y\|_2 = (\sum_{i=1}^n y_i^2)^{1/2}$) of the error

* Originally published in *Izvestiya Vysshchikh Uchebnykh Zavedenii, Radiofizika*, 8 (1966), pp. 1010-1015. Submitted on January 18, 1965. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† S. M. Kirov Ural Polytechnic Institute, Vtuzgorodok, Sverdlovsk 2, USSR.

vector of the terminal state of the system, $\int_0^T v^2(\tau) d\tau$ is the cost of the second player's control [2], and the given number λ is a weighting factor ($\lambda > 0$).

We take it that each player selects his control from his own class of admissible controls. Let U and V be the admissible control classes of the first and second player, respectively. We shall assume that $U \subset W$ is the class of piecewise-continuous functions $u(t)$, $0 \leq t \leq T$, satisfying the condition

$$\|u\| \leq M, \quad M = \text{const.} > 0,$$

while the class V is the whole space $L^2(0, T)$. For any choice $u \in U$ by the first player, Player II selects the control $v_u \in L^2(0, T)$ for which

$$(2) \quad J_2(u, v_u) = \min_{v \in L^2} J_2(u, v).$$

Therefore, Player I must choose a control u^0 such that

$$(3) \quad J_1(u^0, v_{u^0}) = \min_{u \in U} J_1(u, v_u).$$

The control u^0 is optimal for the first player. The optimal control for the second player, $v^0 = v_{u^0}$, is determined from (2) with $u = u^0$, i.e.,

$$J_2(u^0, v^0) = \min_{v \in V} J_2(u^0, v).$$

The problem is to seek the optimal controls for both players.

2. Solution of the problem. The solution of (1) at the terminal instant T is determined by the Cauchy formula

$$(4) \quad \begin{aligned} x(T) = F(T, 0)x_0 + \int_0^T F(T, \tau)b(\tau)u(\tau) d\tau \\ + \int_0^T F(T, \tau)c(\tau)v(\tau) d\tau, \end{aligned}$$

where $\dot{F}(t, t_0) = AF(t, t_0)$, $F(t_0, t_0) = E$, E is the identity matrix. We introduce the notation:

$$(5) \quad \begin{aligned} f = F(T, 0)x_0, \quad l(\tau) = F(T, \tau)b(\tau), \quad h(\tau) = F(T, \tau)c(\tau), \\ Su = \int_0^T l(\tau)u(\tau) d\tau, \quad Pv = \int_0^T h(\tau)v(\tau) d\tau, \quad m_u = f + Su. \end{aligned}$$

Let us assume that a control $u \in U$ has been chosen and let us consider the problem of finding the control v_u which satisfies (2) [2].

Taking (4) and (5) into account, we have

$$(6) \quad J_2(u, v) = \|m_u + Pv\|_2^2 + \lambda \int_0^T v^2(\tau) d\tau.$$

Let the symbol (a, b) denote the scalar product of the elements a and b . If $a, b \in X_n$, then $(a, b) = \sum_{i=1}^n a_i b_i$ and $\|a\|_2^2 = (a, a)$; if $a, b \in L^2(0, T)$, then $(a, b) = \int_0^T a(\tau)b(\tau) d\tau$. In these notations

$$\begin{aligned} J_2(u, v) &= (m_u + Pv, m_u + Pv) + \lambda(v, v) \\ &= (m_u, m_u) + 2(m_u, Pv) + (Pv, Pv) + \lambda(v, v). \end{aligned}$$

For each $v \in L^2(0, T)$ we construct the family of functions $v_\mu = v_u + \mu v$, where μ is a numerical parameter. Then

$$\begin{aligned} J_2(u, v_\mu) &= (m_u, m_u) + (Pv_u, Pv_u) + \lambda(v_u, v_u) + 2(m_u, Pv_u) \\ (7) \quad &+ 2\mu[(m_u, Pv) + (Pv_u, Pv) + \lambda(v_u, v)] \\ &+ \mu^2[(Pv, Pv) + \lambda(v, v)]. \end{aligned}$$

Since the functional $J_2^u(v_\mu) = J_2(u, v_\mu)$ takes its minimal value $J_2^u(v_u)$ when $\mu = 0$, for all $v \in L^2(0, T)$,

$$(8) \quad \left. \frac{dJ_2^u(v_u + \mu v)}{d\mu} \right|_{\mu=0} = 0.$$

Therefore, if we substitute (7) into (8), we obtain

$$(9) \quad (m_u, Pv) + (Pv_u, Pv) + \lambda(v_u, v) \equiv 0.$$

We know that if $a \in X_n$, then $(a, Pv) = (P^*a, v)$, where P^* is the operator adjoint to P . The element P^*a belongs to the space $L^2(0, T)$ and is determined by the equality $[P^*a](\tau) = \sum_{i=1}^n a_i h_i(\tau)$. Thus, from (9) we have

$$(10) \quad (P^*(m_u + Pv_u) + \lambda v_u, v) \equiv 0.$$

Since (10) is valid for $v = P^*(m_u + Pv_u) + \lambda v_u$, and since $(a, a) = 0$ implies $a = 0$, we obtain an integral equation for the unknown function v_u :

$$(11) \quad P^*(m_u + Pv_u) + \lambda v_u = 0.$$

Let $z = Pv_u$. From (11) we have

$$(12) \quad v_u = -\frac{1}{\lambda} P^*(m_u + z),$$

and, therefore,

$$(13) \quad z = -\frac{1}{\lambda} PP^*(m_u + z) = -\frac{1}{\lambda} \int_0^\tau h(\tau)h'(\tau) d\tau(m_u + z),$$

where the prime denotes the transpose. The numerical matrix $H = \int_0^\tau h(\tau)h'(\tau) d\tau$ is nonnegative definite; therefore, when $\lambda > 0$, the matrix $H + \lambda E$ is nonsingular, and from (13) follows

$$z = -(H + \lambda E)^{-1} H m_u.$$

Considering that $[P^*a](\tau) = h'(\tau)a = \sum_{i=1}^n a_i h_i(\tau)$, we finally obtain from (12) that

$$(14) \quad v_u(\tau) = -h'(\tau)(H + \lambda E)^{-1} m_u = -h'(\tau) B m_u,$$

where we have denoted $B = (H + \lambda E)^{-1}$.

The choice of control u by the first player leads the second player to select the control v_u defined by (14). The functional J_1 now takes the value

$$(15) \quad J_1(u, v_u) = \|x(T, u, v_u) - r\|_1 = \|f + Su + P v_u - r\|_1.$$

If for v_u we substitute its value from (14) and carry out the necessary simplifications, in place of (15) we obtain

$$(16) \quad J_1(u, v_u) = \|\lambda B f - r + \lambda B S u\|_1 = \|d + Qu\|_1,$$

where we have introduced the notations: $d = \lambda B f - r$, $Qu = \lambda B S u$. The element d is a known element of the state space X_n , while the linear operator Q maps the element $u \in W$ into the element $Qu \in X_n$.

The problem of minimizing the functional (16) under the constraint $\|u\| \leq M$ has been studied in many papers, for example, in [3]. Let R denote the space of elements x from the state space X_n with the norm $\|x\|_1$ and let R^* be its dual. Then [4]

$$(17) \quad \|x\|_1 = \max_{g \in R^*, \|g\| \leq 1} \langle g, x \rangle,$$

where the symbol $\langle a, b \rangle$ denotes a linear functional on the elements $a \in R^*$, $b \in R$. Taking (17) into account, from (3) and (16) we obtain

$$(18) \quad J_1(u^0, v^0) = \min_{u \in U} J_1(u, v_u) = \min_{u \in U} \max_{g \in R^*, \|g\| \leq 1} \langle g, d + Qu \rangle.$$

The function $\varphi(g, u) = \langle g, d + Qu \rangle$ is linear in both its arguments, and the sets $\|u\| \leq M$ and $\|g\| \leq 1$ are compact, which allows us to apply the minimax theorem [5]:

$$(19) \quad \min_{u \in U} \max_{\|g\| \leq 1} \varphi(g, u) = \max_{\|g\| \leq 1} \min_{u \in U} \varphi(g, u).$$

Since $\min_{\|u\| \leq M} \langle g, d + Qu \rangle = \langle g, d \rangle + \min_{\|u\| \leq M} \langle g, Qu \rangle$, and $\min_{\|u\| \leq M} \langle g, Qu \rangle = \min_{\|u\| \leq M} \langle Q^*g, u \rangle = -\|Q^*g\|M$, (18) and (19) lead to the expression

$$(20) \quad J_1(u^0, v^0) = \max_{\|g\| \leq 1} [\langle g, d \rangle - \|Q^*g\|M].$$

Note that the element Q^*g has the form

$$[Q^*g](\tau) = \lambda g' Bl(\tau) = \lambda \sum_{i=1}^n g_i [Bl(\tau)]_i,$$

and belongs to the space W^* which is the dual of the control space W of the first player. In particular, if $W = L^p(0, T)$, $p > 1$, then $Q^*g \in L^q(0, T)$, $q > 1$, where $1/p + 1/q = 1$. Here

$$\|Q^*g\| = \left(\int_0^T \left| \lambda \sum_{i=1}^n g_i [Bl(\tau)]_i \right|^q d\tau \right)^{1/q}.$$

If $p \rightarrow \infty$, then in the limit

$$\begin{aligned} \|u\| &= \operatorname{ess} \max_{0 \leq \tau \leq T} |u(\tau)|, \\ \|Q^*g\| &= \int_0^T \left| \lambda \sum_{i=1}^n g_i [Bl(\tau)]_i \right| d\tau. \end{aligned}$$

Let $g^\bullet \in R^*$ be the vector which solves problem (20), i.e., is such that

$$\langle g^0, d \rangle - \|Q^*g^0\|M = \max_{g \in R^*; \|g\| \leq 1} [\langle g, d \rangle - \|Q^*g\|M].$$

Then the optimal control u^0 satisfies the relation

$$\langle Q^*g^0, u^0 \rangle = \min_{\|u\| \leq M} \langle Q^*g^0, u \rangle = -\|Q^*g^0\|M.$$

Thus, the optimal control u^0 is an extremal element for the functional Q^*g^0 . If $u \in L^p(0, T)$, then

$$(21) \quad u^0(\tau) = -M \left(\int_0^T |\alpha(t)|^q dt \right)^{-1/p} |\alpha(\tau)|^{q-1} \operatorname{sgn} \alpha(\tau),$$

where we have denoted $\alpha(\tau) = [Q^*g^0](\tau)$. When

$$\|u\| = \operatorname{ess} \max_{0 \leq \tau \leq T} |u(\tau)|,$$

we have

$$(22) \quad u^0(\tau) = -M \operatorname{sgn} \alpha(\tau).$$

If the functional $[\langle g, d \rangle - \|Q^*g\|M]$ is negative for all g , $\|g\| = 1$, then $g^0 = \Theta$, $J_1(u^0, v^0) = 0$. In this case the optimal control is not unique. If everywhere we replace M by m , where the number m , $0 < m < M$,

satisfies the equation $\max_{\|g\|=1} [\langle g, d \rangle - \|Q^*g\|m] = 0$, then the procedure we have described yields the optimal control with the smallest norm m .

Taking (21) or (22) into account, from (14) we obtain an explicit expression for the optimal control v^0 of the second player:

$$v^0(\tau) = -h'(\tau)Bf + h'(\tau)BM \left(\int_0^T |\alpha(t)|^q dt \right)^{-1/p} \int_0^T l(\Theta) |\alpha(\Theta)|^{q-1} \operatorname{sgn} \alpha(\Theta) d\Theta$$

if $1 < p < \infty$, or

$$v^0(\tau) = -h'(\tau)Bf + h'(\tau)BM \int_0^T l(\Theta) \operatorname{sgn} \alpha(\Theta) d\Theta$$

if $p \rightarrow \infty$.

Thus, the problem stated in §1 has been reduced to the variational problem of finding the n -dimensional vector g^0 which minimizes the functional $[\langle g, d \rangle - \|Q^*g\|M]$. This problem is equivalent to that of finding the extremum of a function of n variables g_1, g_2, \dots, g_n ; methods for solving this latter problem have been described, for example, in [6].

In conclusion, we emphasize once again that the assumption that the players may not enter into an agreement is significant in the problem we have considered. In general, an a priori agreement can lead to a coordinated choice of strategies (control functions) so as to yield better results for both players. It is also interesting to note that if the order of priority in selecting the controls is changed, i.e., Player I chooses his control knowing the choice of Player II, the results for Player I may turn out to be worse than in the situation described in this article. This fact, which seems paradoxical at first glance, is a consequence of the nonantagonistic interests of the players.

The author thanks R. Gabasov and Yu. I. Alimov for their discussions on the present work.

REFERENCES

- [1] A. M. LETOV, *Analytical design of controllers*, Automat. Remote Control, 21 (1960), pp. 303-306.
- [2] R. BELLMAN, I. GLICKSBERG AND O. A. GROSS, *Some aspects of the mathematical theory of control processes*, Rept. R-313, The RAND Corporation, Santa Monica, California, 1958.
- [3] R. GABASOV AND F. M. KIRILLOVA, *On a method of solving certain optimal control problems*, Automat. Remote Control, 25 (1964), pp. 289-297.
- [4] B. Z. VULIKH, *Introduction to Functional Analysis for Scientists and Technologists*, Addison-Wesley, Reading, Massachusetts, 1963.
- [5] KY FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42-47. (Russian transl. in *Infinite Antagonistic Games*, N. N. Vorob'ev, ed., Fizmatgiz, Moscow, 1963, pp. 31-39.)
- [6] A. A. FEL'DBAUM, *Computers in Automatic Systems*, Fizmatgiz, Moscow, 1959.

CONVERSE THEOREMS FOR STOCHASTIC LIAPUNOV FUNCTIONS*

H. KUSHNER†

Summary. If a continuous strong Markov process has certain stochastic stability properties, it is proved that a stochastic Liapunov function (see Kushner [1], [2], [3]) exists.¹ This is a stochastic counterpart to a converse theorem of Massera [4] for the deterministic case, where the “process” is the solution to an ordinary differential equation.

1. Nomenclature and some assumptions. Let x_s be a homogeneous strong Markov process and Q a bounded open set, containing the origin, in the state space of the process. Define $t \cap s = \min(t, s)$, and the stopped process $\tilde{x}_t = x_{t \cap \tau}$, where $\tau = \inf\{t: x_t \notin Q\}$. If $x_t \in Q$ for all $t < \infty$, then set $\tau = \infty$. τ is a Markov time of both processes x_t and \tilde{x}_t , [5]; i.e., the event $\{\tau \leq t\}$ is in the σ -algebra determined by conditions on x_s , $s \leq t$, or on \tilde{x}_s , $s \leq t$. If $\tau = \infty$, let x_t be continuous for $t < \infty$ w.p.1; if $\tau < \infty$, let x_t be continuous for $t \leq \tau$ w.p.1. The process \tilde{x}_t is a continuous strong Markov process [5, Theorem 10-2]. Note that $x_\tau \in \partial Q$ w.p.1, relative to $\{\tau < \infty\}$. Write $P_x\{A\}$ and $E_x f$ for the probability of the event A and expectation of the random variable f , *conditioned on the initial value* $x_0 = x$, respectively.

A function $f(x)$ is in the domain of the weak infinitesimal operator ($f(x) \in \mathfrak{D}(\tilde{A}_Q)$) of \tilde{x}_t if, for $x \in Q$,

$$[E_x f(\tilde{x}_t) - f(x)]/t \rightarrow b(x),$$

$$E_x b(\tilde{x}_t) \rightarrow b(x)$$

as $t \rightarrow 0$, for any $x \in Q$, where the limits exist pointwise. We also suppose that if P is open and $P \subset Q$ and $f(x) \in \mathfrak{D}(\tilde{A}_Q)$, then $f(x) \in \mathfrak{D}(\tilde{A}_P)$ (where $f(x)$ is now restricted to P), and on P , $\tilde{A}_Q f(x) = \tilde{A}_P f(x)$. This is not a serious restriction. Denote the weak infinitesimal operator of x_s by \tilde{A} .

Define $I_x(\omega, s, R)$ as the indicator of the (ω, s) set where $\tilde{x}_s(\omega) \in R$ and

* Received by the editors September 8, 1966, and in revised form November 11, 1966.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Aeronautics and Space Administration under Grant NGR 40-002-015, and in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR-693-66.

¹ Prior to this only one converse theorem was available, that of Khasminskii [7], for strong diffusion processes whose differential generators are uniformly elliptic outside all neighborhoods of the origin, and which degenerate in a prescribed way at the origin. This is a special case of our Theorem 2.

$x_0 = x$. Write

$$t_x(\omega, R) = \int_0^\infty I_x(\omega, s, R) ds.$$

$t_x(\omega, R)$ is the *total* time that \tilde{x}_s spends in the measurable set R . Write $|x|$ for any norm of x .

2. Stochastic stability. The following theorem (along with several other homogeneous and nonhomogeneous case theorems) is proved in [1], [2].

THEOREM 1. *Let x_t be a strong Markov process which is continuous up to τ . Let Q be a bounded open set and make the following assumptions:*

- (A1) $V(x) \in \mathfrak{D}(\tilde{A}_Q)$.
- (A2) $V(0) = 0$; $V(x)$ positive definite and continuous in Q .
 $V(\partial Q) = q$; $V(x) < q$, $x \in Q$.
- (A3) $\tilde{A}_Q V(x) = -k(x) \leq 0$ in Q , where $k(x)$ is continuous.
- (A4) $P_x\{\sup_{t \geq s \geq 0} \|x_s - x\| \geq \epsilon > 0\} \rightarrow 0$ as $t \rightarrow 0$ for x in Q .

Then:

- (C1) $P_x\{x_t \rightarrow \partial Q \text{ as } t \rightarrow \tau\} = P_x\{x_t \notin Q, \text{ some } t < \infty\} \leq V(x)/q$.
- (C2) $x_t \rightarrow \{x: x \in Q, k(x) = 0\} = M$, as $t \rightarrow \infty$, with a probability at least $1 - V(x)/q$.
- (C3) Let N_ϵ be an ϵ ($\epsilon > 0$) neighborhood of M . Then $T_x(Q - N_\epsilon) \equiv E_x t_x(\omega, Q - N_\epsilon) < \infty$ and $\sup_{x \in Q} T_x(Q - N_\epsilon) < \infty$ for all $\epsilon > 0$ (in fact, $T_x(\omega, Q - N_\epsilon) \leq V(x)/\delta_\epsilon$, where $\delta_\epsilon = \inf_{x \in Q - N_\epsilon} k(x) > 0$).
- (C4)² $P_x\{\sup_{\infty > t \geq 0} |x_t| \geq \epsilon\} \rightarrow 0$ as $|x| \rightarrow 0$ for all $\epsilon > 0$.

If $k(x)$ is positive definite in a neighborhood of the origin, then

- (C5) $P_x\{\limsup |x_t| > 0\} \rightarrow 0$ as $|x| \rightarrow 0$.

The theorem is proved using a combination of Dynkin's formula [5, Corollary, p. 133] and the super-martingale convergence theorems (applied to $V(\tilde{x}_t)$).

Note that if x_s is the (deterministic) solution to an ordinary differential equation with a nonrandom initial condition, then $V(x)$ is an ordinary Liapunov function. The results (C1) and (C2) are rather useful in the analysis of the asymptotic properties of the random paths of a strong Markov process. If the problem were deterministic, then $V(x)/q$ in (C1), (C2) would be replaced by zero. In the stochastic case, the statements hold with the given probabilities. (C2) and (C3) have applications in stochastic control theory, where it may be desired that a (controlled) process x_s reach some target set at some finite (Markov) time. (See Kushner [2], [6].) (C4) is part of one of the standard definitions of "stability w.p.1" (see [1], [7]).

In the deterministic case, (C1) reduces to Liapunov stability of the origin,

² (C4) implies that the origin is an absorbing state of the process [5].

(C2) reduces to Liapunov asymptotic "stability of the set M ", and (C3) reduces to equiasymptotic "stability of the set M ". (C3) is an indispensable condition which is implied by the existence of a Liapunov function, but is not usually involved in the various definitions of stability (analogous to equiasymptotic stability in the deterministic case).

3. Converse theorems. Three converse theorems are proved. Let M be the origin. Theorem 2 proves the existence of a Liapunov function in some neighborhood of the origin. Under an additional condition, this neighborhood differs from Q by an arbitrarily small volume (Theorem 3). If $x_t \rightarrow 0$ w.p.1 for all initial conditions and another condition holds, then a Liapunov function may be defined on the entire state space (Theorem 4). The proof of Theorem 2 is a stochastic analog of the deterministic proof of Massera [4].

We require the following condition (continuity in probability with respect to the initial condition):

(A5) $P_{x,y}\{|\tilde{x}_t - \tilde{y}_t| \geq \epsilon > 0\} \rightarrow 0$ as $|x - y| \rightarrow 0$ for any $\epsilon > 0$, where x_t and y_t correspond to initial conditions x and y , respectively, in Q .

THEOREM 2. *Assume the conditions of §1, and also (C3), (C4), (A4), (A5), and that M is the origin. Then there exist a function $V(x)$ and an open set $P \subset Q$ containing the origin, such that (A1) to (A3) hold for P replacing Q , and some $p > 0$ replacing q . $k(x)$ is continuous and positive definite in P and $k(0) = 0$.*

Proof. We will actually prove the equivalent, that there is a continuous, bounded and nondecreasing function $c(s)$ satisfying $c(s) > 0$ for $s > 0$ and $c(0) = 0$ so that, for some $B < \infty$,

$$W(x) \equiv E_x \int_0^\tau c(|x_s|) ds \leq B$$

for all $x \in Q$; furthermore $W(x)$ is continuous and is positive definite in a neighborhood of the origin, $W(x) \in \mathfrak{D}(\tilde{A}_Q)$ and $\tilde{A}_Q W(x) = -c(|x|)$.

First we show that *if there is such a $c(s)$, then $W(x)$ has the stated properties.* Let $x \in Q$. Then, by [5, Theorem 3.11] which allows us to write $E_x W(x_{t \cap \tau})$ as $E_x \int_{t \cap \tau}^\tau c(|x_s|) ds$, we obtain

$$E_x W(\tilde{x}_t) - W(x) = -E_x \int_0^{t \cap \tau} c(|x_s|) ds,$$

which, by (A4), equals $-c(|x|)E_x(t \cap \tau) + o(t)$. Now, $E_x(t \cap \tau)/t \rightarrow 1$ as $t \rightarrow 0$ by $P_x\{\tau > 0\} = 1$, (A4), and the monotone convergence theorem. That $E_x c(|\tilde{x}_t|) \rightarrow c(|x|)$ as $t \rightarrow 0$ is easily verified from (A4). Thus $W(x) \in \mathfrak{D}(\tilde{A}_Q)$ and $\tilde{A}_Q W(x) = -c(|x|)$.

(C4) and the method of construction of $c(s)$ may be used to prove that $W(x) \rightarrow 0$ as $|x| \rightarrow 0$. (See below for construction; we have $E_x \int_0^\tau c(|x_s|) ds \leq P_x\{\sup_{\infty > t \geq 0} |x_t| \geq \epsilon\} \max c(s) + \sum_{i: r_i \leq \epsilon} T_i c_i$, which by choosing ϵ small and then $|x|$ small can be made arbitrarily small.)

Note that $E_x \int_{\tau \cap t}^\tau c(|x_s|) ds \rightarrow 0$ as $t \rightarrow \infty$ and that τ depends on x . Let $x \in Q$ and $\delta > 0$ be given. There are an $\epsilon > 0$ and $t < \infty$ so that $|x - y| < \epsilon$ implies³ $E_y \int_{t \cap \tau}^\tau c(|y_s|) ds \leq \delta$. Using this and, by (A5), $E_{x,y} \int_0^t |c(|\tilde{x}_s|) - c(|\tilde{y}_s|)| ds \rightarrow 0$ as $|x - y| \rightarrow 0$, the continuity of $W(x)$ can be proved.

Now, a suitable $c(s)$ will be constructed using (C3). Let N_i be a sequence of spherical neighborhoods of the origin with radii r_i , $r_i \rightarrow 0$ as $i \rightarrow \infty$, and suppose that $N_0 \supset Q$. (Zero is the only accumulation point of $\{r_i\}$.) By (C3),

$$\sup_{x \in Q} T_x(Q - N_i \cap Q) \equiv T_i < \infty$$

for each i . Also T_i is nondecreasing as $i \rightarrow \infty$. Define a bounded sequence $c_i \rightarrow 0$ so that $\sum c_i T_i = B < \infty$. There is obviously a bounded, nondecreasing, continuous, positive definite function $c(s)$ satisfying $c(0) = 0$ and $c(r_i) = c_i$. Now, by the method of construction of $c(s)$ and the definition of the integral, it is easily verified that

$$W(x) = E_x \int_0^\tau c(|x_s|) ds \leq B.$$

Note that if M (in (C3)) is more than the origin, other forms of construction of $c(s)$ can be used.

LEMMA 1. Assume the conditions of Theorem 2. Then $\tau = \infty$ implies that $x_t \rightarrow 0$ w.p.1 (relative to $\tau = \infty$).

Proof. The process x_t spends a finite average time in $Q - N_\epsilon$ for all $\epsilon > 0$. The lemma follows from this, together with the continuity condition (A4) and (C4) (which implies that the trajectory stays in N_ϵ once it gets sufficiently close to the origin, with a high probability).

LEMMA 2. Assume the conditions of Theorem 2. Then

$$P_x\{x_t \rightarrow \partial Q \text{ as } t \rightarrow \tau\} = P_x\{x_\tau \in \partial Q\} \equiv f(x) \in \mathfrak{D}(\tilde{A}_Q)$$

and $\tilde{A}_Q f(x) = 0$.

Proof. The proof is immediate by direct computation, using the evaluation

$$E_x P_{x_t \cap \tau}\{x_\tau \in \partial Q\} = P_x\{x_\tau \in \partial Q\},$$

³ y_s is the process corresponding to the initial condition y .

and noting that, by Lemma 1, either $\tau = \infty$ (in which case $x_t \rightarrow 0$ w.p.1) or $\tau < \infty$ (in which case $x_\tau \in \partial Q$ w.p.1).

Theorem 3 is an extension of a result of Khasminskii [7] and requires only probabilistic (as opposed to probabilistic and partial differential equations) arguments. We require here that the probability that x_t reach ∂Q (before the origin) is a continuous function of the initial condition in Q , and that this probability approaches 1 as $x \rightarrow \partial Q$. This supposition is intended as a slightly more general version of the suppositions of Khasminskii.

Remark. The condition $P_x\{x_t \rightarrow \partial Q \text{ as } t \rightarrow \tau\} \rightarrow 1 \text{ as } x \rightarrow \partial Q$ is satisfied for diffusion processes with differential generators that are elliptic in $Q - N_\epsilon$ for all neighborhoods, N_ϵ , of the origin. We also require that, for initial values in Q , there is a nonzero probability of reaching the origin before ∂Q . The latter condition is trivial, since, if it does not hold in some neighborhood of the origin, there cannot be asymptotic stability.

THEOREM 3. *Assume the conditions of Theorem 2 and that $f(x) \equiv P_x\{x_\tau \in \partial Q\}$ is uniformly continuous in Q and tends to 1 as $x \rightarrow \partial Q$, and $f(x) < 1$ in Q . Then there are a function $V(x)$ and set P which satisfy the conclusions of Theorem 2, and P differs from Q by an arbitrarily small volume.*

Proof. $f(x)$ satisfies the conditions on $V(x)$ (of Theorem 2) in Q , with $p = 1$, except that $\tilde{A}_Q f(x) = 0$. Now, it is clear that, for each $K > 0$, the continuous function $V(x)$ given by

$$\begin{aligned} V(x) &= f(x) + W(x)/K, \\ \tilde{A}_Q V(x) &= -c(|x|)/K \end{aligned}$$

satisfies the statement in Theorem 2 in a set $P_K \subset Q$, with $p_K = 1$. The P_K increase as K increases, and $Q - P_K$ tends to the null set as $K \rightarrow \infty$.

For the global analog of Theorem 2, we need a modification of (C3). Let M equal the origin, again.

(C3') Define $S_r = \{x: |x| < r\}$ and S_r^c as the complement of S_r . Let $\sup_x T_x(S_{r_1} - S_{r_2}) < \infty$ for each $\infty > r_1 > r_2 > 0$, and let $T_x(S_r^c) < \infty$ for all finite x , for any $0 < r < \infty$. $T_x(S_r^c) \rightarrow \infty$ as $|x| \rightarrow \infty$.

THEOREM 4. *Let $x_t \rightarrow 0$ w.p.1 as $t \rightarrow \infty$ for all initial values and let x_t be continuous in $[0, \infty)$. Assume the conditions of §1, and (C3'), (C4), (A4) and (A5). There is a positive definite function $c(s)$ with $c(0) = 0$ and $c(s)$ bounded away from zero outside of any neighborhood of 0, so that*

$$W(x) \equiv \int_0^\infty c(|x_s|) ds < \infty$$

for $|x| < \infty$. Also $W(x) \in \mathcal{D}(\tilde{A})$, $\tilde{A}W(x) = -c(|x|)$, $W(x)$ is positive definite, $W(0) = 0$ and $W(x) \rightarrow \infty$ as $|x| \rightarrow \infty$. $W(x)$ is bounded in any bounded set.

Remark. A stability theorem states that the existence of such a pair $W(x)$, $c(|x|)$ implies (C3') (see Kushner [1], [2]).

Proof. We will construct the function $c(s)$ and show $W(x) \rightarrow \infty$ as $|x| \rightarrow \infty$. The rest of the proof is like that of Theorem 2.

Let $r_i \rightarrow 0$ as in the proof of Theorem 2. Then define

$$T_i \equiv \sup_x T_x(S_{r_i} - S_{r_{i+1}}) < \infty, \quad i > 0,$$

$$T_0(x) \equiv T_x(S_{r_0}^c) < \infty, \quad r_0 > 0.$$

Define a bounded decreasing sequence c_i so that $\sum_1^\infty c_i T_i < \infty$ and define $c_0 = \sup c_i < \infty$. Then there is a bounded continuous, nondecreasing function $c(s)$ with $c(0) = 0$, $c(r_i) = c_i$ and $c(s) = c_0$ for $s \geq r_0$. Also, by construction $W(x) = E_x \int_0^\infty c(|x_s|) ds < \infty$ for all finite x . The last sentence of (C3') implies $W(x) \rightarrow \infty$ as $|x| \rightarrow \infty$.

Remark. If $T_x(S_r^c) \rightarrow m < \infty$ as $x \rightarrow \infty$ along some path, then our $W(x)$ may not tend to infinity. If $W(x)$ is uniformly bounded ($T_x(S_r^c) \leq B < \infty$), then a new radially symmetric $c(s)$ may be defined so that $W(x) \rightarrow \infty$. If $T_x(S_r^c)$ tends to infinity along some paths, and is bounded on others, a non-radially symmetric $c(x_s)$ has to be used. It is desirable that $W(x) \rightarrow \infty$ since such a Liapunov function implies that $x_t \rightarrow 0$ w.p.1 for all initial conditions x . The problem with $T_x(S_r^c)$ bounded along some unbounded path does not appear to be solved for even the deterministic case.

REFERENCES

- [1] H. J. KUSHNER, *On the theory of stochastic stability*, Advances in Control Systems, vol. 4, Academic Press, New York, 1966.
- [2] ———, *Stochastic Stability and Control*, Academic Press, New York, to appear.
- [3] ———, *On the stability of stochastic dynamical systems*, Proc. Nat. Acad. Sci. U. S. A., 53 (1965), pp. 8–12.
- [4] J. L. MASSERA, *On Liapunov's conditions of stability*, Ann. of Math., 50 (1949), pp. 705–721.
- [5] E. B. DYNKIN, *Markov Processes*, Springer-Verlag, Berlin, 1965.
- [6] H. J. KUSHNER, *Stochastic stability and the design of feedback controls*, Proc. Polytechnic Institute of Brooklyn Symposium on Systems Theory, Polytechnic Press, New York, 1965, pp. 177–196.
- [7] R. F. KHASMINSKII, *On the stability of the trajectories of Markov processes*, J. Appl. Math. Mech., 26 (1962), pp. 1554–1565.
- [8] J. L. MASSERA, *Converse theorems of Liapunov's second method*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 158–163.

LIAPUNOV'S DIRECT METHOD AND THE NUMBER OF ZEROS WITH POSITIVE REAL PARTS OF A POLYNOMIAL WITH CONSTANT COMPLEX COEFFICIENTS*

SIEGFRIED H. LEHNIGK†

1. Introduction. To prove by means of Liapunov functions Hermite's stability criterion and Hermite's theorem on the number of zeros with positive real parts of a complex polynomial (i.e., a polynomial with real or nonreal coefficients), we need four theorems of Liapunov's direct method.

THEOREM 1. *The equilibrium of the equation $\dot{x} = Ax$ is asymptotically stable if and only if for an arbitrarily given, positive definite Hermitian matrix C there exists a positive definite Hermitian solution matrix B of $\bar{A}^T B + BA = -C$.*

THEOREM 2. *If the equilibrium of the equation $\dot{x} = Ax$ with $A \neq aU$ (U is the unit matrix, a is a scalar) is asymptotically stable and if R is a singular matrix such that $R(\exp At)x_0 \neq 0$ for every $x_0 \neq 0$, then there exists exactly one positive definite Hermitian solution matrix B of $\bar{A}^T B + BA = -\bar{R}^T R$.*

THEOREM 3. *The equilibrium of the equation $\dot{x} = f(x)$ is asymptotically stable if there exists a function $v(x)$ which is positive definite in $|x| \leq X$ and whose total derivative $\dot{v}(x)$ for the given equation is negative semidefinite in $|x| \leq X$ and of such a nature that $\dot{v}(x(t, x_0, t_0)) \neq 0$ for every nontrivial solution $x(t, x_0, t_0)$ of $\dot{x} = f(x)$ with $|x_0| < X$.*

Proofs of Theorems 1, 2, and 3 for the case that all quantities occurring are real may be found in [2, Th. III 2.2, p. 47; Gen. of Th. III 2.2, p. 51; for the condition $A \neq aU$, see pp. 48–51; Gen. of Th. III 1.2, p. 30]. These theorems remain valid in the complex case, provided that the independent variable, t , is real and that the functions $v(x)$ and $\dot{v}(x)$ have the needed reality properties. These conditions are certainly satisfied for linear equations $\dot{x} = Ax$ with t real. In going from the real to the complex case one has, of course, to replace the transpose of a vector or a matrix by the conjugate complex transpose. (The transpose is denoted by the superscript T and the conjugate complex by an overbar.)

THEOREM 4. *If the equilibrium of the equation $\dot{x} = Ax$ with $A \neq aU$ is completely unstable and if R is a singular matrix such that $R(\exp At)x_0 \neq 0$ for every $x_0 \neq 0$, then there exists exactly one negative definite Hermitian solution matrix B of $\bar{A}^T B + BA = -\bar{R}^T R$.*

Proof. (a) If the equilibrium of $\dot{x} = Ax$ is completely unstable, then all

* Received by the editors August 6, 1966, and in revised form November 21, 1966.

† United States Army Missile Command, Physical Sciences Laboratory, Research and Development Directorate, Redstone Arsenal, Alabama 35809.

eigenvalues of the $n \times n$ matrix A have positive real parts [2, Th. II 3, p. 20]. This implies that, if s_ν , $\nu = 1, \dots, n$, are the not necessarily distinct eigenvalues of the complex matrix A , $\prod_{\mu, \nu=1; \mu \leq \nu}^n (s_\mu + s_\nu) \neq 0$, so that a unique Hermitian solution matrix B of $A^T B + BA = -\bar{R}^T R$ exists and $B \neq 0$ since $\bar{R}^T R \neq 0$ [2, complex version of Th. III 2.1, p. 44].

(b) The matrix B is negative definite. For, if B were not negative definite, the function $-\bar{x}^T Bx$ for $\dot{x} = -Ax$ would not be positive definite. But, according to Theorem 2, $-B$ must be positive definite since the equilibrium of $\dot{x} = -Ax$ is asymptotically stable and since $R(\exp(-At))x_0 \neq 0$ for every $x_0 \neq 0$ as a consequence of the assumption that $R(\exp At)x_0 \neq 0$ for every $x_0 \neq 0$.

2. Two theorems of Hermite. By purely algebraic means, Hermite [1] solved the problem of determining the number of zeros with positive real parts of a complex polynomial without having to compute them numerically. In this paper we will give proofs of Hermite's theorems using Liapunov functions.

With the complex polynomial

$$f(s) = \sum_{\nu=0}^n a_\nu s^{n-\nu},$$

in which, for convenience, we set $a_0 = 1$, we associate the real symmetric $n \times n$ matrix

$$\nabla = \| h_{\mu\nu} \|_{\mu, \nu=1}^n,$$

the Hermite matrix of $f(s)$, with

$$\begin{aligned} h_{\mu\nu} = h_{\nu\mu} &= (\mu - 1, \nu) + (\mu - 2, \nu + 1) + \dots + (1, \mu + \nu - 2) \\ &+ (0, \mu + \nu - 1), \end{aligned} \quad \mu \leq \nu,$$

and

$$\begin{aligned} (\alpha, \beta) &= 0 \quad \text{if } \beta > n, \\ (\alpha, \beta) &= i^{\alpha+\beta+1} ((-1)^\beta \bar{a}_\alpha \bar{a}_\beta + (-1)^{\alpha+1} \bar{a}_\alpha a_\beta), \\ \alpha &= 0, 1, \dots, n-1, \quad \alpha < \beta \leq n. \end{aligned}$$

We need the following principal minors

$$\det \| h_{\mu\nu} \|_{\mu, \nu=1}^\rho = \nabla_\rho, \quad \rho = 1, \dots, n,$$

of the matrix ∇ , the Hermite determinants of $f(s)$. Since ∇ is real, each ∇_ρ is real.

THEOREM 5 (Hermite Criterion [1], [2, Th. IV 2.1, p. 80]). *The complex polynomial $f(s) = \sum_{\nu=0}^n a_\nu s^{n-\nu}$, $a_0 = 1$, is Hurwitzian (i.e., $f(s)$ has no zeros*

with nonnegative real parts) if and only if the numbers $\nabla_1, \dots, \nabla_n$ are positive.

THEOREM 6 ([1], [2, Th. IV 6.2, p. 118]). *Let the numbers $\nabla_\rho, \rho = 1, \dots, n$, for the complex polynomial $f(s) = \sum_{\nu=0}^n a_\nu s^{n-\nu}$, $a_0 = 1$, be different from 0. Then $f(s)$ has $r = \mathcal{V}(1, \nabla_1, \dots, \nabla_n)$ zeros with positive real parts and $n - r$ zeros with negative real parts. ($\mathcal{V}(1, \nabla_1, \dots, \nabla_n)$ is the number of variations in sign contained in the sequence $\{1, \nabla_1, \dots, \nabla_n\}$.)*

Having Theorem 6, one can also deal with the case that not all numbers ∇_ρ for $f(s)$ are different from 0 (see [2, pp. 126–151]).

In [3], Parks gives a proof of Theorem 5 by using a Liapunov function. In this paper (§4) we will prove also Theorem 6 by means of Liapunov functions. *With such Liapunov type proofs available, it is possible to treat every aspect of the stability theory of linear motions by means of the tools provided by Liapunov's direct method.*

3. The Jacobi matrix of $f(s)$ and its Jordan normal form. With the complex polynomial

$$f(s) = \sum_{\nu=0}^n a_\nu s^{n-\nu}, \quad a_0 = 1, \quad n \geq 1,$$

we associate its $n \times n$ Jacobi matrix S , the rows of which are

1st row: $\| -iq_n, 1, 0, \dots, 0 \|$,

μ th row: $\| \underbrace{0, \dots, 0}_{\mu-2}, -p_{n-\mu+2}, -iq_{n-\mu+1}, 1, 0, \dots, 0 \|$, $\mu = 2, \dots, n-1$,

n th row: $\| 0, \dots, 0, -p_2, -p_1 - iq_1 \|$.

The quantities p_ν and q_ν in the elements of S are real and they are defined by the continued fraction expansion of the rational function $g(s)/f(s)$ with

$$(1) \quad g(s) = \frac{1}{2} \sum_{\nu=1}^n (a_\nu + (-1)^{\nu-1} \bar{a}_\nu) s^{n-\nu}.$$

This expansion is formally given by

$$\frac{g(s)}{f(s)} = \frac{p_1}{s + p_1 + iq_1}$$

if $n = 1$, and by

$$(2) \quad \frac{g(s)}{f(s)} = \frac{p_1}{s + p_1 + iq_1 + \frac{p_2}{s + iq_2 + \frac{p_3}{s + iq_3 + \frac{p_4}{s + iq_4 + \frac{p_5}{s + iq_5 + \frac{p_6}{s + iq_6 + \frac{p_7}{s + iq_7 + \frac{p_8}{s + iq_8 + \frac{p_9}{s + iq_9 + \frac{p_{10}}{s + iq_{10}}}}}}}}}}}}}}$$

if $n \geq 2$. (For details concerning this expansion, the reader is referred to Wall [4, Chap. 9, §40; Chap. 10, §45].)

Instead of the matrix S let us consider for a moment that matrix S^* which we obtain from S by reflecting the elements of S in the secondary diagonal. It follows from the definition of the p_r and q_r and from the general recurrence relations for the r th numerator and the r th denominator of a continued fraction that $\det(sU - S^*)$ is the n th denominator of the continued fraction expansion of $g(s)/f(s)$, i.e.,

$$\det(sU - S^*) = f(s).$$

Since S and S^* are related to each other by a unimodular transformation, we also have

$$\det(sU - S) = f(s).$$

This important relation makes it now possible to make statements about the zeros of $f(s)$ by making statements about the stability properties of the equilibrium of the equation

$$\dot{x} = Sx.$$

The matrix S has the essential property of being nonderogatory. If S_κ are the distinct eigenvalues of S of multiplicities π_κ , $\kappa = 1, \dots, k \leq n$, $\sum_{\kappa=1}^k \pi_\kappa = n$, then to each S_κ there corresponds only one linearly independent eigenvector. This follows from the fact that, for every κ , the cofactor of the last element of the first column of $S_\kappa U - S$ is different from 0; it is a lower triangular determinant the diagonal elements of which are 1.

Let M be a matrix which transforms S into its Jordan normal form J ,

$$(3) \quad M^I S M = J = \text{diag} \| J_1, \dots, J_k \|.$$

(The superscript I denotes the inverse.) The J_κ are $\pi_\kappa \times \pi_\kappa$ matrices of the form

$$(4) \quad J_\kappa = S_\kappa U + E_{\pi_\kappa}, \quad \kappa = 1, \dots, k,$$

and the rows of E_{π_κ} are

$$\mu\text{th row: } \| \underbrace{0, \dots, 0}_\mu, 1, \underbrace{0, \dots, 0}_{\pi_\kappa - \mu - 1} \|, \quad \mu = 1, \dots, \pi_\kappa - 1,$$

$$\pi_\kappa\text{th row: } \| 0, \dots, 0 \|.$$

Since S is nonderogatory, the matrix M is, in terms of its column vectors, of the form

$$M = \| m_1, m_2^{(1)}, \dots, m_{\pi_1}^{(1)}, \dots, m_k, m_2^{(k)}, \dots, m_{\pi_k}^{(k)} \|$$

Here m_κ is the linearly independent eigenvector for the eigenvalue S_κ , and the $m_\rho^{(\kappa)}$ are the Jordan vectors for S_κ , $\rho = 2, \dots, \pi_\kappa$, $\kappa = 1, \dots, k$. Let

us assume that the numbers p_ν ($\nu = 1, \dots, n$) are different from 0. Then it follows that in each eigenvector

$$m_\kappa^T = \| m_{1,\alpha+1}, \dots, m_{n,\alpha+1} \|,$$

where

$$\begin{aligned} \alpha &= \pi_1 + \dots + \pi_{\kappa-1} \quad \text{if } \kappa = 1, \dots, k, \\ \alpha &= 0 \quad \text{if } \kappa = 1, \end{aligned}$$

the last component is different from 0,

$$(5) \quad m_{n,\alpha+1} \neq 0.$$

For, if the last component of m_κ were equal to 0, then the last component equation of $(S_\kappa U - S)m_\kappa = 0$ would imply that the next to the last component of m_κ were equal to 0 since $p_2 \neq 0$ by assumption. Working back through the remaining $n - 1$ component equations, we would come out with the trivial solution vector of $(S_\kappa U - S)y = 0$ instead of with the non-vanishing eigenvector m_κ .

The quantities p_ν and q_ν occurring in S are determined by the coefficients a_ν of $f(s)$. The expressions for the q_ν in terms of the a_ν are not needed here as we shall see later. However, we need the expressions for the p_ν in terms of the a_ν . We will show that, if $\nabla_\nu \neq 0$, $\nu = 1, \dots, n$,

$$(6) \quad p_1 = \frac{1}{2} \nabla_1, \quad p_\nu = \frac{\nabla_{\nu-2} \nabla_\nu}{\nabla_{\nu-1}^2}, \quad \nu = 2, \dots, n, \quad \nabla_0 = 1,$$

where the ∇_ν are the Hermite determinants of $f(s)$. According to their definition (§2), the ∇_ν are determined by the coefficients a_ν of the polynomial $f(s)$.

To verify (6), let us return to the continued fraction expansion of $g(s)/f(s)$ with $g(s)$ given by (1), and let us set

$$r_1(s) = f(s) - g(s), \quad r_2(s) = g(s).$$

From $r_1(s)$ and $r_2(s)$ we obtain the sequence of identities

$$(7) \quad r_\kappa(s) \equiv (c_\kappa s + k_\kappa) r_{\kappa+1}(s) - r_{\kappa+2}(s), \quad \kappa = 1, \dots, n, \quad r_{n+2}(s) \equiv 0$$

(Euclid's chain of $r_1(s)$ and $r_2(s)$) in which all $r_\kappa(s)$ are polynomials,

$$r_\kappa(s) = \sum_{\nu=0}^{n-\kappa+1} \gamma_{\nu\kappa} s^{n-\nu-\kappa+1}, \quad \kappa = 1, \dots, n+1,$$

where

$$\begin{aligned} \gamma_{\nu 1} &= \frac{1}{2}(a_\nu + (-1)^\nu \bar{a}_\nu), & \nu &= 0, 1, \dots, n, \\ \gamma_{\nu 2} &= \frac{1}{2}(a_{\nu+1} + (-1)^\nu \bar{a}_{\nu+1}), & \nu &= 0, 1, \dots, n-1. \end{aligned}$$

The $r_{\kappa}(s)$ are obtained, of course, under the assumption that the leading coefficients $\gamma_{0\kappa}$ of the $r_{\kappa}(s)$ are different from 0.

The coefficients c_{κ} in (7) are determined by

$$(8) \quad c_{\kappa} = \frac{\gamma_{0\kappa}}{\gamma_{0,\kappa+1}}, \quad \kappa = 1, \dots, n,$$

and between the numbers c_{κ} and the numbers p_{ν} in (2) we have the relations

$$(9) \quad \frac{1}{c_1} = p_1, \quad \frac{1}{c_{\kappa-1} c_{\kappa}} = -p_{\kappa}, \quad \kappa = 2, \dots, n.$$

With the coefficients $\gamma_{\nu 1}$ and $\gamma_{\nu 2}$ of the polynomials $r_1(s)$ and $r_2(s)$ we form the following $(2\rho - 1) \times (2\rho - 1)$ determinants:

$$(10) \quad P_{\rho} = \det \begin{vmatrix} \gamma_{02} & \gamma_{12} & \gamma_{22} & \gamma_{32} & \cdots & \cdot \\ \gamma_{01} & \gamma_{11} & \gamma_{21} & \gamma_{31} & \cdots & \cdot \\ 0 & \gamma_{02} & \gamma_{12} & \gamma_{22} & \cdots & \cdot \\ 0 & \gamma_{01} & \gamma_{11} & \gamma_{21} & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \gamma_{\rho-1,2} \end{vmatrix}$$

$$\rho = 1, \dots, n, \quad \gamma_{\nu\kappa} = 0 \text{ if } \nu > n - \kappa + 1, \quad \kappa = 1, 2.$$

We immediately see that

$$P_1 = \gamma_{02} = p_1,$$

if we observe (8) and (9), and that $\gamma_{01} = 1$ since $a_0 = 1$. Furthermore, one can show that

$$(11) \quad P_{\rho} = \gamma_{0,\rho+1} \prod_{\nu=2}^{\rho} \gamma_{0\nu}^2, \quad \rho = 2, \dots, n,$$

(see [4, Chap. 9, §41, Proof of Th. 41.1, pp. 165–166]). With (8) and (9), from (11) we obtain

$$P_{\rho} = (-1)^{\rho-1} \prod_{\nu=1}^{\rho} p_{\nu} P_{\rho-1}, \quad \rho = 1, \dots, n, \quad P_0 = 1,$$

and thus

$$(12) \quad p_1 = P_1, \quad p_{\nu} = -\frac{P_{\nu-2} P_{\nu}}{P_{\nu-1}^2}, \quad \nu = 2, \dots, n, \quad P_0 = 1.$$

To obtain (6) from these relations, it is convenient to use the Bilharz determinants B_{ρ} of $f(s)$ which can be defined as follows:

$$(13) \quad B_{\rho} = 2^{\rho} Q_{\rho}, \quad \rho = 1, \dots, n,$$

where Q_ρ is the following $(2\rho - 1) \times (2\rho - 1)$ determinant:

$$(14) \quad Q_\rho = \det \begin{vmatrix} \gamma_1' & \gamma_2'' & -\gamma_3' & -\gamma_4'' & \gamma_5' & \gamma_6'' & \cdots & \cdot \\ 1 & \gamma_1'' & -\gamma_2' & -\gamma_3'' & \gamma_4' & \gamma_5'' & \cdots & \cdot \\ 0 & \gamma_1' & \gamma_2'' & -\gamma_3' & -\gamma_4'' & \gamma_5' & \cdots & \cdot \\ 0 & 1 & \gamma_1'' & -\gamma_2' & -\gamma_3'' & \gamma_4' & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots & (-1)^\vartheta \kappa_\rho \end{vmatrix}$$

$$\rho = 1, \cdots, n, \quad \gamma_\nu' = \gamma_\nu'' = 0 \text{ if } \nu > n,$$

in which

$$\gamma_\nu' = (\bar{a}_\nu + a_\nu)/2, \quad \gamma_\nu'' = i(\bar{a}_\nu - a_\nu)/2,$$

$$\gamma_\nu' = \gamma_\nu'' = 0 \quad \text{if } \nu > n,$$

$$\kappa_\rho = \begin{cases} \gamma_\rho' & \text{if } \rho \text{ is odd,} \\ \gamma_\rho'' & \text{if } \rho \text{ is even,} \end{cases}$$

$$\vartheta = \left[\frac{\rho + 1}{2} \right] + 1,$$

where, x being real, $[x]$ is the largest integer $\leq x$.

In terms of the coefficients $a_\nu = \gamma_\nu' + i\gamma_\nu''$ of $f(s)$, $a_0 = 1$, the coefficients $\gamma_{\nu 1}$ of $r_1(s) = f(s) - g(s)$ are

$$\gamma_{\nu 1} = \gamma_\nu', \quad \nu = 0, 2, 4, \cdots,$$

$$\gamma_{\nu 1} = i\gamma_\nu'', \quad \nu = 1, 3, 5, \cdots;$$

and the coefficients $\gamma_{\nu 2}$ of $r_2(s) = g(s)$ are

$$\gamma_{\nu 2} = \gamma_{\nu+1}', \quad \nu = 0, 2, 4, \cdots,$$

$$\gamma_{\nu 2} = i\gamma_{\nu+1}'', \quad \nu = 1, 3, 5, \cdots.$$

With these relations, from (10) we obtain

$$(15) \quad P_\rho = \det \begin{vmatrix} \gamma_1' & i\gamma_2'' & \gamma_3' & i\gamma_4'' & \cdots & \cdot \\ 1 & i\gamma_1'' & \gamma_2' & i\gamma_3'' & \cdots & \cdot \\ 0 & \gamma_1' & i\gamma_2'' & \gamma_3' & \cdots & \cdot \\ 0 & 1 & i\gamma_1'' & \gamma_2' & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \lambda_\rho \end{vmatrix}$$

$$\rho = 1, \cdots, n, \quad \gamma_\nu' = \gamma_\nu'' = 0 \text{ if } \nu > n,$$

with

$$\lambda_\rho = \begin{cases} \gamma_\rho' & \text{if } \rho \text{ is odd,} \\ i\gamma_\rho'' & \text{if } \rho \text{ is even.} \end{cases}$$

If we multiply now in the determinants (15) rows and columns by factors i and -1 such that all elements of the new determinants are real and such that the first nonvanishing element in each row is positive, we see that

$$P_\rho = (-1)^{\rho(\rho-1)/2} Q_\rho, \quad \rho = 1, \dots, n,$$

with Q_ρ given by (14), and that, if we observe (13),

$$P_\rho = (-1)^{\rho(\rho-1)/2} 2^{-\rho} B_\rho, \quad \rho = 1, \dots, n.$$

From (12) it follows then that

$$p_1 = \frac{1}{2} B_1, \quad p_\nu = \frac{B_{\nu-2} B_\nu}{B_{\nu-1}^2}, \quad \nu = 2, \dots, n, \quad B_0 = 1.$$

Since $B_\nu = \nabla_\rho$, $\nu = 1, \dots, n$, (see [2, Chap. 4 §3]), relations (6) follow at once.

4. Proofs of Theorems 5 and 6 by means of Liapunov functions. As a Liapunov function for the equation

$$\dot{x} = Sx$$

let us use the Hermitian form

$$v(x) = \bar{x}^T Bx,$$

with

$$B = \text{diag} \left\| \frac{\nabla_n}{\nabla_{n-1}}, \frac{\nabla_{n-1}}{\nabla_{n-2}}, \dots, \frac{\nabla_2}{\nabla_1}, \nabla_1 \right\|$$

under the assumption that $\nabla_\rho \neq 0$, $\rho = 1, \dots, n$. Forming the total derivative of $v(x)$ for $\dot{x} = Sx$, we obtain the matrix equation

$$(16) \quad \bar{S}^T B + BS = -C.$$

According to the particular structure of S and B and according to the first of the relations (6) it turns out that

$$(17) \quad C = \text{diag} \parallel 0, \dots, 0, \nabla_1^2 \parallel.$$

All numbers q_ν in the elements of S drop out in forming C . This is the reason that they need not be known in detail.

Using the matrices M and J introduced in §3, we may replace (16) by the equation

$$(18) \quad \bar{J}^T B^* + B^* J = -C^*,$$

in which

$$B^* = \bar{M}^T B M = \parallel b_{\mu\nu}^* \parallel, \quad C^* = \bar{M}^T C M = \parallel c_{\mu\nu}^* \parallel$$

are Hermitian matrices. Since C is of the special form (17), the matrix C^* has in its main diagonal in particular the nonvanishing elements

$$(19) \quad c_{\alpha+1, \alpha+1}^* = |\nabla_1|^2 |m_{n, \alpha+1}|^2 \neq 0,$$

where

$$\begin{aligned} \alpha &= \pi_1 + \cdots + \pi_{\kappa-1} \quad \text{if } \kappa = 1, \cdots, k, \\ \alpha &= 0 \quad \text{if } \kappa = 1. \end{aligned}$$

These elements are different from 0 since $\nabla_1 \neq 0$ and because of the inequalities (5) which in turn are a consequence of the assumption that each ∇_ρ , $\rho = 1, \cdots, n$, is different from 0.

It is now possible to see that the assumption $\nabla_\rho \neq 0$, $\rho = 1, \cdots, n$, implies that the matrix S has no eigenvalues with vanishing real parts, i.e., that the polynomial $f(s)$ has no zeros with vanishing real parts if $\nabla_\rho \neq 0$. Suppose that $\nabla_\rho \neq 0$, $\rho = 1, \cdots, n$, and that S has at least one eigenvalue with vanishing real part. Without loss of generality we may assume that $S_1 = i\Omega_1$, Ω_1 real, is one of them. Then, if we observe (3) and (4), (18) shows us that

$$-i\Omega_1 b_{11}^* + i\Omega_1 b_{11}^* = -c_{11}^* = 0$$

in contradiction to (19) for $\kappa = 1$.

Proof of Theorem 5. Necessity. Suppose that $f(s) = \sum_{\nu=0}^n a_\nu s^{n-\nu}$, $a_\nu = \gamma'_\nu + i\gamma''_\nu$, $a_0 = \gamma'_0 = 1$, is Hurwitzian. Then the numbers p_ν , $\nu = 1, \cdots, n$, in the elements of the matrix S are different from 0. This follows from the facts that the p_ν are the partial numerators of the continued fraction expansion of $g(s)/f(s)$ with $g(s) = \frac{1}{2}(f(s) + (-1)^{n-1}\bar{f}(-s))$, $\bar{f}(s) = \sum_{\nu=0}^n \bar{a}_\nu s^{n-\nu}$, $\bar{a}_0 = a_0 = 1$, and that, if the zeros of $f(s)$ have negative real parts, the zeros of $g(s)$ have vanishing real parts so that the polynomials $g(s)$ and $f(s)$ are relatively prime (see [4, Chap. 10, §§45, 46, 47]). This implies that the Hermite determinants ∇_ρ , $\rho = 1, \cdots, n$, of $f(s)$ are different from 0. In particular, we have $2p_1 = \nabla_1 = (0, 1) = 2\gamma'_1 > 0$ since γ'_1 is the negative sum of the real parts of the zeros of $f(s)$. This proves the necessity of the condition of Theorem 5 for $n = 1$.

If $n \geq 2$, the matrix $-C$ given by (17) is strictly negative semidefinite and the matrix S in $\dot{x} = Sx$ is not a scalar multiple of the unit matrix. Furthermore, if $x(t, x_0, t_0)$ is a solution of $\dot{x} = Sx$ with $x_0 \neq 0$, we have

$$\bar{x}^T(t, x_0, t_0) C x(t, x_0, t_0) = |\nabla_1|^2 |x_n(t, x_0, t_0)|^2 \neq 0.$$

For, if this function were identically equal to 0, we would have $x_n(t, x_0, t_0) \equiv 0$. This would imply either $x(t, x_0, t_0) \equiv 0$ or $p_\nu = 0$ for at least one $\nu \geq 2$. But we have $x_0 \neq 0$ and $p_\nu \neq 0$, $\nu = 1, \cdots, n$.

Now, let

$$R = \text{diag} \parallel 0, \dots, 0, \nabla_1 \parallel.$$

Then $C = \tilde{R}^T R$. Hence, using the general form of the solutions of $\dot{x} = Sx$, $x(t, x_0, t_0) = (\exp S(t - t_0))x_0$, we have

$$R(\exp St)x_0 \neq 0 \quad \text{for every } x_0 \neq 0.$$

Consequently, by Theorem 2, the matrix

$$B = \text{diag} \parallel \frac{\nabla_n}{\nabla_{n-1}}, \frac{\nabla_{n-1}}{\nabla_{n-2}}, \dots, \frac{\nabla_2}{\nabla_1}, \nabla_1 \parallel$$

is positive definite, i.e., the numbers ∇_ρ , $\rho = 1, \dots, n$, are positive.

Sufficiency. Suppose that the Hermite determinants ∇_ρ , $\rho = 1, \dots, n$, of $f(s)$ are positive. Then, if $n = 1$, $-\tilde{x}^T Cx$ and $\tilde{x}^T Bx$ reduce to negative and positive scalars, respectively. Hence, by Theorem 1, the equilibrium of $\dot{x} = Sx$ is asymptotically stable, i.e., the zero of $f(s)$ is negative. If $n \geq 2$, $-\tilde{x}^T Cx$ is a strictly negative semidefinite form which does not vanish identically along every nontrivial solution of $\dot{x} = Sx$ and the matrix B is positive definite. Consequently, by Theorem 3, the equilibrium of $\dot{x} = Sx$ is asymptotically stable, i.e., $f(s)$ is Hurwitzian.

Proof of Theorem 6. Let the Hermite determinants ∇_ρ , $\rho = 1, \dots, n$, of $f(s)$ be different from 0 and let $\mathfrak{U}(1, \nabla_1, \dots, \nabla_n) = r > 0$. Then the sequence $\{\nabla_1, \nabla_2/\nabla_1, \dots, \nabla_n/\nabla_{n-1}\}$ contains r negative elements and $n - r$ positive elements.

(a) Suppose that S has $0 \leq r' < r$ eigenvalues with positive real parts, i.e., suppose that S has $n - r' > n - r$ eigenvalues with negative real parts. (Remember that there are no eigenvalues with vanishing real parts.) Without loss of generality we may assume that S_1, \dots, S_l are the distinct eigenvalues of S with negative real parts with multiplicities π_1, \dots, π_l , $\pi_1 + \dots + \pi_l = n - r'$. To these eigenvalues there corresponds the following submatrix J' of the Jordan normal form J of S ,

$$(20) \quad J' = \text{diag} \parallel J_1, \dots, J_l \parallel.$$

Let

$$B' = \parallel b_{\mu\nu}^* \parallel_{\mu,\nu=1}^{n-r'} \quad \text{and} \quad C' = \parallel c_{\mu\nu}^* \parallel_{\mu,\nu=1}^{n-r'}$$

be the corresponding submatrices of B^* and C^* , respectively. We consider the matrix relation

$$(21) \quad \bar{J}'^T B' + B' J' = -C',$$

with J' given by (20). The matrix $-C'$ is strictly negative semidefinite (or $-C'$ is a negative scalar if $n - r' = \pi_1 = 1$, see (19)) since $-C$ and $-C^*$

have this property. Furthermore, if $\dot{x}'(t, x_0', t_0)$ with $x_0' \neq 0$ is a solution of the equation $\dot{x}' = J'x'$, the equilibrium of which is asymptotically stable, we have

$$\bar{x}'^T(t, x_0', t_0)C'x'(t, x_0', t_0) \neq 0$$

as a consequence of the particular structure of C' and J' (see (19) and (21)). Consequently, since

$$\prod_{\substack{\mu, \nu=1 \\ \mu \leq \nu}}^{n-r'} (s_\mu + s_\nu) \neq 0$$

($s_1, \dots, s_{n-r'}$ are the not necessarily distinct eigenvalues of J'), a unique Hermitian solution B' of (21) exists. If $n - r' > 1$, J' is not proportional to the unit matrix. Hence, by Theorem 2, B' is positive definite. This implies that the sequence $\{\nabla_1, \nabla_2/\nabla_1, \dots, \nabla_n/\nabla_{n-1}\}$ contains at least $n - r' > n - r$ positive elements, a contradiction to the fact that it contains only $n - r$ positive elements. If $n - r' = 1$, i.e., if $r = n$, $-C'$ reduces to a negative scalar, i.e., $-C'$ is negative definite. Hence, by Theorem 1, the scalar B' is positive, i.e., the sequence $\{\nabla_1, \nabla_2/\nabla_1, \dots, \nabla_n/\nabla_{n-1}\}$ contains at least one positive element, a contradiction to the fact that it contains no positive elements. Hence, r' cannot be less than r . We have $r' \geq r$.

(b) If $r = n$, our result shows that $r' = r = n$, which proves Theorem 6 for $r = n$. Suppose now that $0 < r < n$ and that S has $r' > r$ eigenvalues with positive real parts. We proceed then as in part (a) of this proof, using the same notations as there, however now with S_1, \dots, S_l being the eigenvalues with positive real parts and $\pi_1 + \dots + \pi_l = r'$. The equilibrium of $\dot{x}' = J'x'$ is then completely unstable. By Theorem 4, the uniquely determined matrix B' in (21) is negative definite, i.e., the sequence $\{\nabla_1, \nabla_2/\nabla_1, \dots, \nabla_n/\nabla_{n-1}\}$ contains $r' > r$ negative elements in contradiction to the fact that it contains only r negative elements. Consequently, $r' = r$.

REFERENCES

- [1] CH. HERMITE, *Sur le nombre des racines d'une équation algébrique comprises entre des limites données*, J. Reine Angew. Math., 52 (1856), pp. 39-51.
- [2] S. H. LEHNIGK, *Stability Theorems for Linear Motions with an Introduction to Liapunov's Direct Method*, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.
- [3] P. C. PARKS, *Comment on 'The frequency domain solution of regulator problems'*, IEEE Trans. Automatic Control, AC-11 (1966), p. 344.
- [4] H. S. WALL, *Analytical Theory of Continued Fractions*, Van Nostrand, New York, 1948.

SURVEILLANCE PROBLEMS: TWO-DIMENSIONAL WITH CONTINUOUS SURVEILLANCE*

G. R. ANTELMAN, C. B. RUSSELL AND I. R. SAVAGE†

Abstract. This paper is concerned with optimal strategies for controlling a two-component Poisson production process under continuous surveillance. Either or both components may be repaired at any time at some cost. A form of long run average income is used as the criterion for comparing strategies. The class of stationary Markovian strategies which do not permit hesitation and which allow production on only a finite number of states is shown to be a complete class of strategies under a certain assumption of monotonicity on the income function.

Properties of optimal strategies are derived and the results applied to a numerical example.

1. Introduction. In this paper we will be concerned with a two-component production process which can be described at any specific time by a pair $z = (x, y)$ of nonnegative integers or by the type of repair which is in progress. Inspections are costless at any time, so we speak of continuous surveillance. Repair costs, rates of income from production, and the stochastic nature of the process will be described below. Our aim is to obtain a strategy which will maximize a form of the long run average income.

We first examine the class of all possible strategies for the control of the production process. General theorems are proved which eliminate strategies involving hesitation and those containing an infinite number of feasible production states (§3). Then we are able to apply a theorem of Derman [1] which yields the class of admissible strategies. Finally, we determine qualitative properties of the optimal strategy (§4–§7) and apply the results to a numerical example (§8).

The model discussed here is analogous to the continuous surveillance model used by Savage [3]. We replace his one-dimensional process by one of two dimensions. Higher dimensions cause no new difficulties. Part of the work here on elimination of strategies places the work of Savage [3] on a sounder theoretical basis (see Theorem 3.5). Theorems 3.5 and 4.2 show that Derman's [1] linear programming analysis would be applicable in looking for the optimal policy.

A glossary of the symbols used is provided at the end of the paper.

2. Model for continuous surveillance. Three types of repair will be considered in the sequel:

R_0 : repairing both components,

* Received by the editors August 1, 1966, and in revised form November 15, 1966.

† Department of Statistics, Florida State University, Tallahassee, Florida. This work was supported by the Office of Naval Research under Contract Nonr 988(13).

R_x : repairing the x -component,

R_y : repairing the y -component.

The repair $R_0(R_x)[R_y]$ requires a positive number $m_0(m_x)[m_y]$ of units of time at a cost of $K_0(K_x)[K_y]$ per unit of time. After entering a repair state r_0, r_x or r_y from (x, y) , the process must remain in the repair state for the specified time. When this time has elapsed, the process will enter state $(0, 0)$ if R_0 has been performed, $(0, y)$ if R_x , and $(x, 0)$ if R_y .

Let $z(t)$ represent the state of the production process at time t . We shall assume that at time 0 the process has just entered the repair state r_0 . The process will remain in r_0 for the required m_0 time units before moving to the production state $(0, 0)$. Thus $z(m_0) = (0, 0)$. While in the production state z , the income per unit of time will be denoted by $i(z)$.

During production the process will be a Markov chain with probability p of moving from (x, y) to $(x + 1, y)$ and probability q of moving from (x, y) to $(x, y + 1)$, where $p \geq 0, q \geq 0$ and $p + q = 1$. The times between transitions will be independently exponentially distributed with unit parameter. Thus, the probability of waiting more than T units of time for a transition is given by e^{-T} .

Let S denote a strategy, that is, a rule specifying precisely when to begin repairs of each kind given the entire history of the process. Before describing how to choose S , we first describe the quantity which we wish to maximize. As t increases $z(t)$ will, with probability 1, move through a sequence of states, the first two elements in that sequence being r_0 and $(0, 0)$. Let $z_0, z_1, \dots, z_n, \dots$ denote that sequence. Let T_n be the time spent in z_n and C_n be the income obtained while in z_n . Notice that ET_n and EC_n depend on S . Define

$$(2.1) \quad I(N | S) = \frac{\sum_0^N EC_n}{\sum_0^N ET_n}.$$

Although $ET_n = 0$ is possible, we are saved the embarrassment of division by 0 since it is assumed that $ET_0 = T_0 = m_0 > 0$. Then define

$$(2.2) \quad I(S) = \liminf_{N \rightarrow \infty} I(N | S)$$

and

$$(2.3) \quad I^* = \sup_s I(S).$$

A rule S^* is said to be optimal if $I(S^*) = I^*$. We prove the existence of optimal rules and find some of their properties. Other possible definitions

of optimal could have been considered; for example, if $I(T | S)$ was the expected income per unit of time up to T with rule S , one might seek a rule which would yield $\sup_S (\liminf_{T \rightarrow \infty} I(T | S))$. It is conjectured that this objective is equivalent to the one described above but we have found our first definition tractable. Other definitions using discounting could be considered.

3. Admissible strategies. The theorems in this section will be general results which help to characterize the class of admissible strategies. Assumptions restricting the function $i(z)$ will be held to a minimum.

THEOREM 3.1. *If $i(z)$ is replaced by $i'(z)$, where $i(z) \geq i'(z)$ for each z , or K_0, K_x, K_y are replaced by K'_0, K'_x, K'_y , where $K_0 \leq K'_0, K_x \leq K'_x, K_y \leq K'_y$, then for every N and S the value of $I(N | S)$ will not be increased.*

Proof. The indicated changes only affect the value of the numerator of $I(N | S)$, and the expectations in the numerator will not be increased since the new random variables are stochastically smaller than the original ones.

At this point the class of conceivable strategies is large. In particular we acknowledge the existence of strategies which allow repairs to begin after a stay of positive duration in a production state. A strategy will be said to involve *hesitation* if it allows this to occur with positive probability. The following theorem permits us to exclude such strategies from consideration.

THEOREM 3.2. *If S is a strategy which involves hesitation, then there exists a strategy S' not involving hesitation such that*

$$I(N | S) = I(N | S')$$

for all N . If an optimal strategy exists, then there exists an optimal strategy not involving hesitation.

Proof. It suffices to prove the first statement in the theorem. Let H_n denote the collection of histories of the process up to the entering of the n th state which allow hesitation with positive probability during the n th state under strategy S , and let \bar{H}_n be the collection of those n -stage histories which do not permit hesitation at the n th state. As before, T_n and C_n are random variables denoting the time and income in the n th state. Let n_1 be the index of the state where hesitation can first occur with positive probability. (Notice that if n_1 does not exist then hesitation will not occur, for if we let p_n be the probability that hesitation occurs first at the n th visited state, we have $\Pr(\text{hesitation}) = \sum_{n=1}^{\infty} p_n = \sum_{n=1}^{\infty} 0 = 0$.) Then we can write

$$(3.1) \quad ET_{n_1} = E(T_{n_1} | H_{n_1}) \Pr(H_{n_1}) + E(T_{n_1} | \bar{H}_{n_1}) \Pr(\bar{H}_{n_1})$$

and

$$(3.2) \quad EC_{n_1} = E(C_{n_1} | H_{n_1}) \Pr(H_{n_1}) + E(C_{n_1} | \bar{H}_{n_1}) \Pr(\bar{H}_{n_1}).$$

We shall modify S so that hesitation will not occur at n_1 and so that none of the values of ET_n and EC_n will be changed. If \bar{H}_{n_1} has occurred, no modification of S is necessary. Now consider the case when H_{n_1} occurs. Under strategy S let P_0, P_x, P_y be the probabilities, given H_{n_1} , of beginning repairs of types R_0, R_x, R_y , respectively, before a transition occurs, and let P be the probability given H_{n_1} of a transition before repairs begin.

Now modify S in the following manner: when the n_1 th state is entered and hesitation is allowed (H_{n_1}) under S , immediately begin the various repairs with probabilities P_0, P_x, P_y , or with probability P allow production until a transition occurs. With this new strategy, S' , the location of the process at the beginning of the $(n_1 + 1)$ st stage will have the same distribution as it had under the original strategy S . The expected time and income in the n_1 th state will be shown to be the same for both strategies.

For the strategy S' , given H_{n_1} , the expected time in n_1 is clearly P . For the strategy S let $G(t)$ be the probability that repairs will have begun by time t , given H_{n_1} , and that a transition has not occurred first. Since hesitation can occur, we have $G(0) < 1$ and $\lim_{t \rightarrow \infty} G(t) > 0$. Then, by definition of P , we have

$$(3.3) \quad P = \int_0^{\infty} e^{-t}(1 - G(t)) dt,$$

and with strategy S we compute

$$\begin{aligned} E(T_{n_1} | H_{n_1}) &= \int_0^{\infty} te^{-t} dG(t) + \int_0^{\infty} te^{-t}(1 - G(t)) dt \\ &= te^{-t}G(t) \Big|_0^{\infty} - \int_0^{\infty} G(t)(e^{-t} - te^{-t}) dt \\ &\quad + \int_0^{\infty} te^{-t}(1 - G(t)) dt \\ &= - \int_0^{\infty} G(t)e^{-t} dt + \int_0^{\infty} G(t)te^{-t} dt \\ &\quad + 1 - \int_0^{\infty} te^{-t}G(t) dt \\ &= 1 - \int_0^{\infty} G(t)e^{-t} dt = \int_0^{\infty} e^{-t}(1 - G(t)) dt = P. \end{aligned} \tag{3.4}$$

Since $i(z)$ is constant in any given state, the expected income from a given state is a constant multiple of the expected time in that state. Hence the expected incomes for the two strategies also agree. Thus at the end of the n_1 th stage we have $I(n_1 | S) = I(n_1 | S')$.

As mentioned previously, the distribution of the $(n_1 + 1)$ st state has

not been changed. The two changes which have been made are (a) the elimination of hesitation at the n_1 th stage, and (b) the elimination of $(n_1 + 1)$ -stage histories which involve hesitation. We will therefore keep a synthetic history as follows: upon entering the $(n_1 + 1)$ st state under S' , create by randomization the time which would have been spent in the n_1 th if the strategy S had been used given (a) the true history of the process up to the entering of the n_1 th state and (b) which state was entered at the $(n_1 + 1)$ st stage.

Thus, on entering the $(n_1 + 1)$ st state, the distribution of histories when using strategy S' is the same as when using strategy S , and we may proceed under strategy S until n_2 , the second index where hesitation can occur, is reached. At this stage we modify S in the same way as before.

The proof is not by induction; all of the modifications could be made at one time. We proceeded in this way in an effort to clarify the way that the synthetic histories insure that $I(N | S')$ behaves like $I(N | S)$. These histories are a part of this proof, but will not be required in practice, for Theorem 3.5 eliminates randomization in optimal strategies.

THEOREM 3.3.

$$I^* \geq -\min(K_0, K_x, K_y).$$

Proof. The strategy which keeps the production process in the least costly repair state at all times will yield (exactly) the lower bound for I^* given above.

We will now make assumptions on the form of $i(z)$.

Assumption M. Let $mK = \max(m_0K_0, m_xK_x, m_yK_y)$ and $M = \min(m_0, m_x, m_y, 1)$. We will assume that $i(z)$ is monotone nonincreasing in each coordinate until $i(z) < -mK/M$, and once $i(z) < -mK/M$ it never again reaches this level as x or y increases.

Assumption M implies

$$(3.5) \quad C_0 = \left\{ z: i(z) \geq -\frac{mK}{M} \right\}$$

is a finite set.

*Assumption M**. $i(z)$ satisfies the assumption M and is strictly decreasing in each coordinate until $i(z) < -mK/M$.

THEOREM 3.4. If $i(z)$ satisfies Assumption M, then for each strategy S there exists a strategy S' with the following properties:

(a) $I(S') \geq I(S)$ and

(b) with S' production occurs in only a finite number of states.

If an optimal strategy exists, then there is an optimal strategy satisfying (b).

Proof. Again only the first statement requires proof. We can assume without loss of generality that $i(0, 0) \leq 0$ and that strategy S does not

involve hesitation. First we shall describe a new strategy S' and then show that it has the desired properties.

(1) Let S' be the same as S until a point in \bar{C}_0 , the complement of C_0 , is entered.

(2) Then conduct (with no real time) a random experiment simulating the path of the production process under S until a repair is made or until it is ascertained that no repair will ever be made.

(3) If no repair is ever to be made as a result of the experiment, place the process in r_0 and keep it there forever.

(4) If a repair is to be made, immediately make the repair prescribed by the experiment and repeat that repair as many times as there were states visited by the random experiment until under this one-one correspondence the experiment dictates that the repair would have been completed.

(5) When (4) has been completed, the process is on the same coordinate axis under each strategy. Under S' , however, the position of the process is to the left of or below (possibly, the same as) the position under S . We now begin production using the strategy dictated by the position of the process under S , the past history of the process given by the true history up until \bar{C}_0 was entered, and the synthetic history given by the random experiment until completion of repair. Continue as in step (1).

Condition (b) is satisfied by S' since Assumption M implies C_0 contains only a finite number of points.

With strategy S' , whenever production occurs, the rate of income will be at least as high as with strategy S . This follows from (5) above.

Let us now consider $I(N | S)$ and $I(N | S')$ in detail. We may write

$$(3.6) \quad I(N | S) = \frac{B_1 + B_2}{A_1 + A_2},$$

where, in the first N stages,

B_1 is the expected income from production while in C_0 minus the expected costs of repairs;

B_2 is the expected income from production while in \bar{C}_0 ;

A_1 is the expected amount of production time while in C_0 plus the expected time of repairs; and

A_2 is the expected amount of time in \bar{C}_0 .

In a similar manner for S' let

$$(3.7) \quad I(N | S') = \frac{B_1' + B_2'}{A_1' + A_2'},$$

where the primed quantities are analogous to the unprimed quantities but correspond to the position of the process under S and not under S' . Then

the following relationships hold:

- (a) $A_1 = A_1' > 0$,
- (b) $B_1 \leq B_1' \leq 0$,
- (c) $B_2 \leq B_2' \leq 0$,
- (d) $A_2' = m^* A_2 \geq 0$, where

$$\min(m_0, m_x, m_y) \leq m^* \leq \max(m_0, m_x, m_y),$$

- (e) $m^* B_2 \leq B_2' \leq 0$.

Relations (a) through (d) are clear upon a moment's reflection, while relation (e) can be verified through the following argument:

$$m^* B_2 \leq -m^* \cdot mK \cdot \frac{A_2}{M} = -mK \cdot \frac{A_2'}{M} \leq -\max(K_0, K_x, K_y) \cdot A_2' \leq B_2'.$$

We now proceed to show that $I(N | S') \geq I(N | S)$. To do so we must consider two cases.

Case 1. $m^* \geq 1$. Here relation (d) implies $A_2' \geq A_2 \geq 0$, which yields

$$I(N | S) = \frac{B_1 + B_2}{A_1 + A_2} \leq \frac{B_1' + B_2'}{A_1' + A_2'} = I(N | S')$$

as desired.

Case 2. $m^* < 1$. Since

$$\frac{B_1 + m^* B_2}{A_1 + m^* A_2} \leq \frac{B_1' + B_2'}{A_1' + A_2'},$$

it suffices to show that

$$\frac{B_1 + B_2}{A_1 + A_2} \leq \frac{B_1 + m^* B_2}{A_1 + m^* A_2}.$$

To demonstrate the inequality above, we first notice that the relationship

$$\frac{B_2}{A_2} \leq \frac{B_1}{A_1}$$

is true. Multiplying both sides by the nonnegative quantity $(1 - m^*)A_1A_2$, adding $A_1B_1 + m^*A_2B_2$, and rearranging terms yield

$$A_1B_1 + m^*A_2B_2 + A_1B_2 + m^*A_2B_1 \leq A_1B_1 + m^*A_2B_2 + m^*A_1B_2 + A_2B_1$$

which, when rewritten as

$$(A_1 + m^*A_2)(B_1 + B_2) \leq (A_1 + A_2)(B_1 + m^*B_2),$$

is easily recognizable as the desired result.

THEOREM 3.5. *If $i(z)$ satisfies Assumption M, then there exists an optimal*

stationary nonrandomized strategy for continuous surveillance not using hesitation and with production on at most a finite number of states.

Proof. Theorem 3.2 asserts that hesitation need not be considered and Theorem 3.4 insures that only a finite number of production states are required. The resulting model with time in a state interpreted as a cost is that of Derman [1]. His Theorem 3 is equivalent to the above property. (In the statement of Derman's Theorem 3, "assumption A" is superfluous [2].)

COROLLARY 3.6. *When using an optimal strategy of the kind described in Theorem 3.5, the lattice of pairs of nonnegative integers is partitioned into sets C , R_0 , R_x , R_y , and these sets have the following properties:*

(i) *when the process enters C , it is allowed to continue in production until another state is reached;*

(ii) *when the production process reaches a point in $R_0(R_x)[R_y]$, the repair of both components (the x -component) [the y -component] begins immediately.*

An optimal strategy is equivalent to such a partition. Furthermore C is finite, and the sequence of states visited by the production process using such an optimal strategy is a Markov chain.

The symbol r_0 will be used in the two following ways.

1. It will represent the state where both coordinates are being repaired.

2. It will represent a point $z = (x, y)$ whose entry is the signal for repairs of both coordinates to begin.

The symbol R_0 will represent the set of pairs of nonnegative integers such that repair R_0 begins whenever the set is entered (Corollary 3.6). The context should make the specific usage clear. Similar usages will apply to r_x , r_y , R_x and R_y .

In the following, whenever we refer to an optimal strategy, we shall mean an optimal strategy of the kind described in Theorem 3.5 and Corollary 3.6. Other kinds of strategies need no further consideration except as technical devices in proofs.

4. Details of admissible strategies.

THEOREM 4.1. *If $i(z)$ satisfies Assumption M and $i(0, 0) < -\min(K_0, K_x, K_y)$, then C is empty, $I^* = -\min(K_0, K_x, K_y)$, and the optimal strategy consists of keeping the production process in the least costly repair state.*

Notice that $(0, 0)$ is an element of C unless C is empty. For if $(0, 0)$ is not an element of C , no other production states can be reached. However, $(0, 0)$ could be a transient state of the resulting Markov chain. This Markov chain might have several ergodic classes, although if $(0, 0)$ is a recurrent state there is but one ergodic class.

THEOREM 4.2. *If $i(z)$ satisfies Assumption M, an optimal strategy exists with $(0, 0)$ a recurrent state.*

Proof. If C is empty, then $(0, 0)$ is recurrent although no time is spent there. Therefore, assume that C is not empty and $(0, 0)$ is transient. With probability 1 we eventually reach a point on one of the coordinate axes in $R_0 \cup R_x \cup R_y$. At that point we must make repairs. The resulting repairs will, when completed, find us either at $(0, 0)$ or at the same position we were in when the repairs began. If we are at the same position, we must begin repairs again. Thus the value of the income will be the same as for a strategy always calling for repair. This is true for both coordinate axes. Hence $(0, 0)$ can be made a recurrent event in an optimal strategy.

With the above facts in mind we see that the sequence of states of the production process will be a stationary Markov chain with a finite number of states and one ergodic class. This will be referred to as the embedded chain of the process. It is then clear that for any rule under consideration we can write

$$(4.1) \quad I(S) = \frac{\sum i(z)p_s(z)}{\sum t(z)p_s(z)},$$

where the following hold.

(1) The summation is over the states which can possibly occur when S is used.

(2) $p_s(z)$ is the steady state probability, for the embedded chain under S , of being in state z , which can be either a production state or a repair state. When there is no possible ambiguity, the S will be suppressed.

(3) $i(z)$ has the previous meaning when z is a production state. Otherwise it is the cost of repair. For instance, if z is the repair state r_0 , then $i(z) = -m_0K_0$. When z is a state from which repairs begin, $i(z) = 0$.

(4) $t(z)$ is the expected time in a particular state. For a state in C it is 1, for a state which leads to repairs it is 0, and for the various repair states it is m_0 , m_x , or m_y .

The above representation of $I(S)$ is contained in Derman [1, Theorem 3]. The evaluation of $I(S)$ is equivalent to finding the steady state probabilities of the embedded chain. Another computational method is based on the concept of cycles. A cycle consists of the (random) sequence of visited states starting from $(0, 0)$ until this state is about to be visited again. Notice that $(0, 0)$ can be entered only from a repair state. Thus all cycles are of positive length. Also $(0, 0)$ appears precisely once in each cycle. If we define $n_s(z)$ as the expected number of visits under S to the state z in each cycle, it is easily seen that $n_s(0, 0) = 1$ and that in general $n_s(z) = p_s(z)/p_s(0, 0)$. Thus, an equivalent expression for $I(S)$ as given in (4.1) is

$$(4.2) \quad I(S) = \frac{\sum i(z)n_s(z)}{\sum t(z)n_s(z)}.$$

Now we must find the expected number of visits per cycle instead of the steady state probabilities of the embedded chain.

5. Properties of optimal strategies under the specialized model, $K_x = K_y = \infty$. At this point we are ready to begin the detailed discussion of properties of the optimal strategies. To do this we will make specialized assumptions regarding the function $i(z)$ as well as the other cost parameters of the model. The first special case to be considered was originally examined by Taneja [4]. Assume K_x and K_y are infinitely large, or, in other words, the only kind of repair to be considered is R_0 in which both coordinates are repaired simultaneously. Under this specialized model we can write for each $z = (x, y) \in C$,

$$(5.1) \quad n_s(z) = N_s(z)p^x q^y,$$

where $N_s(z)$ is the number of random walk paths from $(0, 0)$ to (x, y) passing through points of C . Also, for each $z = (x, y) \in C$, we say that C is *full* with respect to z if $z' = (x', y') \in C$ whenever $x' \leq x$ and $y' \leq y$. If C is full with respect to z we can write

$$(5.2) \quad N_s(z) = \binom{x+y}{x}.$$

Again we will suppress the subscript S in $N_s(z)$ and $n_s(z)$ when there is no possibility of ambiguity. We then say that C is full provided that C is full with respect to each of its elements.

By our Assumption M for $i(z)$, the sets $C_\pi = \{z: i(z) \geq \pi\}$ are full for all $\pi \geq -mK/M$. If we consider the partitions of the lattice of pairs of nonnegative integers induced by these sets, we see that under the specialized model there is a natural correspondence between (continuation) sets C_π and strategies S_π . The following theorem informs us that the optimal strategy S is precisely that determined by the set C_{I^*} .

THEOREM 5.1. *If C is the continuation set of an optimal strategy and $i(z)$ satisfies Assumption M, then C is full and*

$$C = \{z: i(z) \geq I^*\}.$$

Proof. Since S is optimal,

$$(5.3) \quad I^* = I(S) = \frac{\sum N(z)p^x q^y i(z) - m_0 K_0}{\sum N(z)p^x q^y + m_0},$$

or

$$(5.4) \quad \sum N(z)p^x q^y [i(z) - I^*] - m_0(K_0 + I^*) = 0.$$

Now we shall show that C must be full and that furthermore I^* plays the role of π . In particular, if there is an element z in C such that $i(z) < I^*$,

then eliminate such an element from C to obtain a new strategy. Then (5.4) will be replaced by an inequality with a $>$ sign in it, which would imply that in (5.3) the $=$ would be replaced by $<$, which would contradict the assumption that C was optimal. Likewise if there is a z not in C such that $i(z) > I^*$, by adding such a z to C to obtain a new strategy we would obtain a $>$ in (5.4) and $<$ in (5.3).

THEOREM 5.2. *Let $i(z)$ satisfy Assumption M. If $\pi_1 > \pi_2 > \pi_3$ and $I(S_{\pi_1}) > I(S_{\pi_2})$, then $I(S_{\pi_2}) \geq I(S_{\pi_3})$. Dually, if $\pi_1 > \pi_2 > \pi_3$ and $I(S_{\pi_2}) < I(S_{\pi_3})$, then $I(S_{\pi_1}) \leq I(S_{\pi_2})$.*

The proof of this theorem requires the following lemma which is easily verified.

LEMMA 5.3. *If $D > a > 0$, $a' > 0$, and $b \geq b'$, then $(J - ab)/(D - a) \geq J/D$ implies that $J/D \geq (J + a'b')/(D + a')$.*

Proof of Theorem 5.2. The inequality $\pi_1 > \pi_2 > \pi_3$ implies $C_{\pi_1} \subset C_{\pi_2} \subset C_{\pi_3}$ since $C_{\pi} = \{z: i(z) \geq \pi\}$. Let S_1, S_2, S_3 be the strategies corresponding to $C_{\pi_1}, C_{\pi_2}, C_{\pi_3}$, respectively, and let $I_c(S_1), I_c(S_2), I_c(S_3)$, $T_c(S_1), T_c(S_2), T_c(S_3)$ be the total expected income during a cycle and the expected duration of a cycle under the designated strategies. Since $C_{\pi_1} \subset C_{\pi_2}$, we have $T_c(S_1) \leq T_c(S_2)$. Similarly $T_c(S_2) \leq T_c(S_3)$. For $i = 1, 2, 3$, $T_c(S_i) > 0$ since, by assumption, repairs take time. Let $T_c(S_2) = D$, $T_c(S_1) = D - a$ and $T_c(S_3) = D + a'$, where $a \geq 0$, $a' \geq 0$. Denote $I_c(S_2) = J$, $I_c(S_1) = J - ab$, $I_c(S_3) = J + a'b'$. We note that $a > 0$ by virtue of $I(S_1) > I(S_2)$. (This is just abbreviated notation for $I(S_{\pi_1}) > I(S_{\pi_2})$ which is our assumption.) Hence we have two cases to consider, namely, $a' = 0$ and $a' > 0$. The first case, $a' = 0$, gives us immediately $I(S_2) = I(S_3)$. In the second case, $a' > 0$, we note that $I(S_1) > I(S_2)$ can be written as

$$(5.5) \quad I(S_1) = \frac{J - ab}{D - a} > \frac{J}{D} = I(S_2).$$

Thus, Lemma 5.3 informs us that

$$(5.6) \quad I(S_2) = \frac{J}{D} \geq \frac{J + a'b'}{D + a'} = I(S_3),$$

provided that $b \geq b'$. This last fact is easily seen as soon as we note that b and b' are the income "rates" in the sets $C_{\pi_2} - C_{\pi_1}$ and $C_{\pi_3} - C_{\pi_2}$ respectively. Result (5.6), then, is the desired conclusion of the theorem.

6. Properties of optimal strategies under the general model. We now return to the general discussion where either or both components can be repaired. With strategy S (§§3 and 4) use the following definitions:

(a) Let $P_c(z, S)$ be the steady state probability that in a cycle the state

z will be visited. Those states z such that $P_c(z, S) > 0$ will be called positive states.

(b) Let $I_c(z, S)$ and $T_c(z, S)$ be the expected income and expected time in a cycle until the positive state z is entered for the first time or until the cycle is over. Denote by $I_c(S | z)$ and $T_c(S | z)$ the expected income and expected time remaining in a cycle after first entering the positive state z .

Then, for any positive state z ,

$$I_c(S) = I_c(z, S) + P_c(z, S)I_c(S | z)$$

and

$$T_c(S) = T_c(z, S) + P_c(z, S)T_c(S | z).$$

If S is optimal, then

$$I^* = I(S) = \frac{I_c(S)}{T_c(S)}$$

or

$$I_c(S) - I^*T_c(S) = 0$$

or

$$(6.1) \quad I_c(z, S) - I^*T_c(z, S) + P_c(z, S)[I_c(S | z) - I^*T_c(S | z)] = 0.$$

The quantity $I_c(S | z) - I^*T_c(S | z)$, which will be denoted by $F_s(z)$, is defined for each positive state z of each strategy S . It will be used extensively below. Notice that the origin will be a positive state for all optimal strategies. Thus, for an optimal S , one has $F_s(0, 0) = 0$. If z is a positive state for an optimal strategy S , then for an arbitrary strategy S' , we have $F_s(z) \geq F_{s'}(z)$. Furthermore, any S such that $F_s(z) = \max_{s'} F_{s'}(z)$ for each positive z under S is optimal. Henceforth the following convention will be used in selecting an optimal S .

(a) If an optimal strategy could be obtained with either $z \in C$ or $z \in R_0 \cup R_x \cup R_y$, then by convention $z \in R_0 \cup R_x \cup R_y$.

(b) If an optimal strategy could be obtained with either $z \in R_x \cup R_y$ or $z \in R_0$, then by convention $z \in R_0$.

(c) If an optimal strategy could be obtained with either $z \in R_x$ or $z \in R_y$, then by convention $z \in R_x$.

THEOREM 6.1. *If S is optimal and $i(z)$ satisfies Assumption M, then for positive states $z \in C \cup R_0 \cup R_x \cup R_y$, $F_s(z)$ is nonincreasing in each coordinate. If S is optimal and $i(z)$ satisfies Assumption M*, then for positive states $z \in C$, $F_s(z)$ is strictly decreasing in each coordinate.*

Proof. The proof will be by contradiction. Assume $z = (x, y)$ and $z' = (x', y')$, with $x \leq x'$ and $y \leq y'$, are positive states satisfying $F_s(z) < F_s(z')$.

Consider a new strategy, S' , defined in the following way.

In each cycle S' is the same as S until the positive state z is entered. Then, until a repair is made, when $z(t) = z''$ use the actions dictated by S when $z(t) = z'' + z' - z$. If the first repair performed is of both components, the cycle is complete, and the definition of S' is finished. If the first repair is made on the x -component, then after the first repair, when $z(t) = z'' = (x'', y'')$ use the action dictated by S when $z(t) = (x'', y'') + (0, y' - y)$. Then continue in this manner until the y -component is repaired. After the y -component is repaired, S' is again the same as S . If the y -component is repaired first, the procedure is homologous.

For any strategy S' , the left-hand side of (6.1) can never be positive. Presently we will show that $F_{S'}(z) \geq F_S(z')$, but then, since S is optimal, we have, from (6.1),

$$\begin{aligned} 0 &= I_c(z, S) - I^*T_c(z, S) + P_c(z, S)F_S(z) \\ &< I_c(z, S') - I^*T_c(z, S') + P_c(z, S')F_S(z') \\ &\leq I_c(z, S') - I^*T_c(z, S') + P_c(z, S')F_{S'}(z). \end{aligned}$$

The first inequality is strict since $P_c(z, S) = P_c(z, S') > 0$ and $F_S(z') > F_S(z)$, giving a contradiction.

We now complete the proof by showing that $F_{S'}(z) \geq F_S(z')$. For each sample path originating at z' consider a (congruent) path originating at z such that at any time until a repair begins, the difference between these two paths is precisely $z' - z$. We continue on the path from z' until S has caused both components to be repaired. If the x -component is repaired first, then the two paths under consideration will differ by $y' - y$ in the second coordinate until the y -component is repaired. If the y -component is repaired first, the pair of paths is homologous, differing by $x' - x$ in the first coordinate.

This pairing of paths has the following two interesting properties.

1. At each instant of time until both components are repaired, the income on the path from z' does not exceed that on the path from z . If $i(z)$ is strictly decreasing, the income from the z path will be strictly greater than that from the z' path until both components are repaired.

2. If A' is a measurable set of paths originating at z' , then $P(A') = P(A)$ where A is the corresponding set originating at z .

These properties clearly imply $F_{S'}(z) \geq F_S(z')$, and, if Assumption M^* is satisfied, $F_{S'}(z) > F_S(z')$, thus completing the proof.

THEOREM 6.2. *If $i(z) < I^*$ and the function $i(z)$ satisfies Assumption M, then $z \notin C$ for an optimal strategy.*

Proof. If z is in C then write $F_S(z)$ in the form

$$F_S(z) = F'_S(z) + F''_S(z),$$

where $F'_s(z)$ is the expected income less I^* per unit of time after entering z and until the first repair is completed, and $F''_s(z)$ is the expected income less I^* per unit of time after leaving z and completing the first repair until the end of the cycle. Note that $F'_s(z)$ is negative since by assumption $i(z) < I^*$, $i(z') < I^*$ until repairs begin, and $-\min(K_0, K_x, K_y) \leq I^*$ from Theorem 3.3.

Consider a strategy S' where S' is the same as S until z is reached, and at that time repairs are begun according to the following randomization procedure: under S find the probabilities of repairing both, the first and the second, components when repairs are begun after leaving z ; use these to generate the decision under S' as to which repairs to begin upon entering z .

The proof will be completed by showing $F_s(z) < F_{s'}(z)$. Define $F'_{s'}(z)$ and $F''_{s'}(z)$ analogously with $F'_s(z)$ and $F''_s(z)$. Then $F'_{s'}(z) > F'_s(z)$ since the expected contributions to $F'_{s'}(z)$ and $F'_s(z)$ are the same for repairs and the contribution to $F'_{s'}(z)$ for production is 0 but the contribution to $F'_s(z)$ from production is negative.

Finally we must show that $F''_{s'}(z) \geq F''_s(z)$. There are no contributions to these terms when both components are repaired, and this happens with the same probability under S and S' . The x -component (y -component) is repaired with the same probability under both S and S' . When the x -component (y -component) is repaired under S' , the $y(x)$ coordinate will never exceed what it would have been under S . So Theorem 6.1 implies $F''_{s'}(z) \geq F''_s(z)$.

Thus $F_s(z) < F_{s'}(z)$ so S is not optimal.

THEOREM 6.3. *When S is an optimal strategy and $i(z)$ satisfies Assumption M, then:*

(i) *If $z = (x, y) \in R_0$ and $z_1 = (x_1, y_1) \in R_x$, then $y_1 < y$. (The set where both components are repaired, R_0 , is above the set where the first component is repaired, R_x .)*

(ii) *If $z = (x, y) \in R_0$ and $z_2 = (x_2, y_2) \in R_y$, then $x_2 < x$. (R_0 is to the right of R_y .)*

(iii) *If $z_1 = (x_1, y_1) \in R_x$ and $z_2 = (x_2, y_2) \in R_y$, then either $y_2 > y_1$ or $x_1 > x_2$. (No element of R_x is to the left and above any element of R_y .)*

(iv) *If $z_1 = (x_1, y_1) \in R_x$, then for any positive integer n , $z = (x_1 + n, y_1) \notin C$. (No element of C is directly to the right of any element of R_x .)*

(v) *If $z_2 = (x_2, y_2) \in R_y$, then for any positive integer n , $z = (x_2, y_2 + n) \notin C$. (No element of C is directly above any element of R_y .)*

(vi) *If $z = (x, y) \in R_0$ and m and n are nonnegative integers, then $z^* = (x + m, y + n) \notin C$. (No element of C is above and to the right of any element of R_0 .)*

(vii) *If $z = (x, y) \in R_0 \cup R_x \cup R_y$, then either $(x + 1, y)$ or $(x, y + 1)$ is an element of $R_0 \cup R_x \cup R_y$.*

Proof. Each conclusion of the theorem is intuitively appealing. It should be noted, however, that this theorem does not imply that C is full. Proofs will be given for (i), (iii) and (iv). The proof of (ii) is analogous to that of (i), the proofs of (v) and (vi) to that of (iv), and (vii) is clear from conclusions (iv), (v) and (vi).

To prove (i) assume the contrary, that is, $z = (x, y) \in R_0$, $z_1 = (x_1, y_1) \in R_x$ and $y_1 \geq y$. Then $F_s(z) = -m_0(K_0 + I^*)$ and $F_s(z_1) = -m_x(K_x + I^*) + F_s(0, y_1)$. Since $y_1 \geq y$, from Theorem 6.1 we have

$$(6.2) \quad F_s(z_1) \leq -m_x(K_x + I^*) + F_s(0, y).$$

(Although $(0, y)$ might not be a positive state for S , the meaning of $F_s(0, y)$ will be clear and the application of Theorem 6.1 can be justified.) Also, $F_s(z) \leq F_s(z_1)$ since, at z_1 , we could repair both components. But

$$(6.3) \quad F_s(z) \geq -m_x(K_x + I^*) + F_s(0, y),$$

since it is not advantageous to repair only the x -component at z . Combining (6.2) and (6.3) one obtains $F_s(z) \geq F_s(z_1)$. Hence $F_s(z) = F_s(z_1)$ and, by convention, $z_1 \in R_0$.

To prove (iii), assume the contrary, that is, $z_1 = (x_1, y_1) \in R_x$, $z_2 = (x_2, y_2) \in R_y$, $y_1 \geq y_2$ and $x_2 \geq x_1$. The following then hold:

$$(6.4) \quad \begin{aligned} F_s(z_1) &= -m_x(K_x + I^*) + F_s(0, y_1) \\ &\leq -m_x(K_x + I^*) + F_s(0, y_2), \end{aligned}$$

$$(6.5) \quad \begin{aligned} F_s(z_2) &= -m_y(K_y + I^*) + F_s(x_2, 0) \\ &\leq -m_y(K_y + I^*) + F_s(x_1, 0), \end{aligned}$$

$$(6.6) \quad F_s(z_1) \geq -m_y(K_y + I^*) + F_s(x_1, 0),$$

$$(6.7) \quad F_s(z_2) \geq -m_x(K_x + I^*) + F_s(0, y_2).$$

Inequalities (6.4) and (6.5) are from Theorem 6.1, and (6.6) and (6.7) result from considering nonoptimal strategies. Then $F_s(z_2) = F_s(z_1)$, which would contradict the convention that z_2 should be in R_x if there is a choice between R_x and R_y .

To prove (iv), again assume the contrary, that is, $z_1 = (x_1, y_1) \in R_x$ and $z = (x_1 + n, y_1) \in C$. From Theorem 6.1, $F_s(z_1) \geq F_s(z)$. At z , the x -component could be repaired so that $F_s(z) \geq F_s(z_1)$. Hence $F_s(z_1) = F_s(z)$, which contradicts the convention that would have placed z in R_x .

THEOREM 6.4. *If z is a positive state in C , then z can be reached from $(0, 0)$ without repairs.*

Proof. Let $z = (x, y)$, and consider the path determined by the sequence of states $(x, y - 1)$, $(x, y - 2)$, \dots , $(x, 0)$. Continue (downward) along this path as long as possible while remaining in positive production states.

Let $z_1 = (x_1, y_1)$ be the last state in this sequence that we visit. Then proceed to the left from z_1 along the path $(x_1 - 1, y_1), (x_1 - 2, y_1), \dots, (0, y_1)$ as far as we are able while remaining in positive production states. Let $z_2 = (x_2, y_2)$ be the last such state along this path. Proceed downward from z_2 , determining $z_3 = (x_3, y_3)$ and continue as above. This procedure will either lead us to the origin, as desired, or to a positive production state $z_0 = (x_0, y_0)$ for which $(x_0 - 1, y_0)$ and $(x_0, y_0 - 1)$ are not positive production states. But then there is no way of reaching z_0 , so the latter alternative is impossible.

7. Results depending on symmetry assumptions. We consider now some special results dependent on the following symmetry assumptions.

Symmetry Assumptions:

- (A) $i(x, y) = i(y, x)$ and $i(x, y)$ satisfies Assumption M*;
- (B) $i(0, 0) = 0$;
- (C) $K_x = K_y \equiv K^*, m_x = m_y \equiv m^*$;
- (D) $p = q = \frac{1}{2}$, i.e., the production process has an equal chance of changing in either coordinate.

THEOREM 7.1. *Under the symmetry assumptions one has, for an optimal strategy S , $F_S(x, y) = F_S(y, x)$; if (x, y) is in R_0 , then (y, x) is in R_0 ; if $(x, y) \in R_x$ and $x \neq y$, then $(y, x) \in R_y$.*

Proof. The opportunities at (x, y) are the same as those at (y, x) .

COROLLARY 7.2. *Under the symmetry assumptions, an optimal strategy S has no points of C either directly above or directly to the right of a point $(x, x) \in R_0 \cup R_x$. There is exactly one point of $R_0 \cup R_x$ on the 45° line $\{(x, x)\}$.*

THEOREM 7.3. *If $i(x, y) = i(x + y)$ and the symmetry assumptions hold, then for a fixed value of $x + y$ and optimal S , one has that $F_S(x, y)$ is a non-decreasing function of $|x - y|$.*

Proof. Let z and z' be two points on the same 135° line. Let the smaller of the two coordinates of z be smaller than the smaller of the two coordinates of z' . This implies that the absolute difference of the coordinates of z is larger than the absolute difference of the coordinates of z' . Assume the contrary, that is, $F_S(z) < F_S(z')$.

Now, as in the proof of Theorem 6.1, consider analogous paths originating at z and z' . Follow these paths until either the one from z' reaches $R_0 \cup R_x \cup R_y$ or they have the same value for their smaller coordinate. Up to that time the income for each path has been the same. If the path from z' has entered $R_0 \cup R_x \cup R_y$, the income remaining for that path will not be greater than the income from the path starting at z , since the path from z will be at least as close to a coordinate axis as the z' -path, and a repair could be made on the z -path. If the smaller coordinates of the two paths are now equal, then the larger coordinates will also be equal, since at each instant of time the two paths must be on the same 135° line. Then the in-

come remaining will be the same by Theorem 7.1. The optimal procedure has been followed from z' but not necessarily from z . However, the income from the z -path has been at least as large as the income from the z' -path. This then yields the conclusion.

8. Example. The following will illustrate the methods of computation discussed in §4. The example is of particular interest in that the assumptions of Theorem 6.3 and §7 are satisfied but the optimal C is not full.

Let $p = q = \frac{1}{2}$, $i(x, y) = -(x + y)$, $m_0 = 5$, $m_x = m_y = 3$, $K_0 = K_x = K_y = 5$. Notice, $i(z)$ is both concave and convex. It will be shown that the C associated with the optimal strategy is not full by evaluating the expected income per unit time for each of a complete class of strategies. As before, let S denote a strategy and $I(S)$ the expected income per unit time for the strategy S . Let S^* be an optimal strategy and C^* its continuation set. First, we give several results concerning inadmissible strategies for this problem.

LEMMA 8.1. *For all $z = (x, y) \in C^*$, $x + y \leq 3$.*

Proof. From Theorem 6.2, $i(z) \geq I^*$ for all $z \in C^*$. It is easily shown (see, for example, S_2 below) that $I^* > -4$.

LEMMA 8.2. *C^* is not empty and the x -axis (y -axis) boundary point of C^* is an element of R_x (R_y).*

Proof. See S_1 and S_2 below.

Strategies will be specified by such configurations as

$$\begin{array}{cccc} r_y & & & \\ c & & r & \\ & r_x & & r \\ c & c & c & r_x. \end{array}$$

The letters are placed at the nonnegative integer lattice points and c means continue, r means repair both components, r_x means repair the x -component, and r_y means repair the y -component. For this example,

$$C = \{(0, 0), (0, 1), (1, 0), (0, 2), (2, 0)\},$$

$$R_y = \{(0, 3)\},$$

$$R_x = \{(3, 0), (1, 1)\},$$

$$R_0 = \{(1, 2), (2, 1)\}.$$

Notice that the present strategy cannot be optimal by (i) of Theorem 6.3. In particular, a better strategy would be obtained by placing $(1, 2)$ in R_y and $(2, 1)$ in R_x .

Before listing the S 's and associated $I(S)$'s for a complete class of S 's, we will introduce some notation and give two examples of the calculation of $I(S)$.

Example 1. Consider the strategy

r_y	u	u	u
c	r	u	u
c	r	r	u
c	c	c	r_x ,

where unreachable (nonpositive) points with $\max(x, y) \leq 3$ have been labelled " u ". Let $N = [n_{jk}]$ represent the matrix of expected number of visits per cycle to the point (j, k) . It will be convenient to label the lowest row " 0 " and the leftmost column " 0 ". Then for the present example it is easy to show that

$$N = \frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 4 & 1 & 0 \\ 8 & 4 & 2 & 1 \end{bmatrix}.$$

If (j, k) is reachable, let $d_{jk}(t_{jk})$ be the contribution to the expected income (time) for the cycle whenever (j, k) is reached. Let $d_{jk}(t_{jk}) = 0$ if (j, k) is unreachable. Then $D = [d_{jk}]$ and $T = [t_{jk}]$ are given by

$$D = - \begin{bmatrix} 15 & 0 & 0 & 0 \\ 2 & 25 & 0 & 0 \\ 1 & 25 & 25 & 0 \\ 0 & 1 & 2 & 15 \end{bmatrix}, \quad T = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 5 & 0 & 0 \\ 1 & 5 & 5 & 0 \\ 1 & 1 & 1 & 3 \end{bmatrix}.$$

The expected income per unit of time I is then given by

$$I = \frac{N \cdot D}{N \cdot T} = -\frac{196}{56} = -3.5,$$

where $N \cdot D \equiv \sum_{j,k} n_{jk} d_{jk}$.

*Example 2.*¹ Consider the strategy

u	r_y	u	u
r_y	c	r_x	u
c	c	c	r_x
c	c	r_x	u .

The D and T matrices for this policy are given by

¹ In this example and in the remainder of §8 we have not followed the convention of §6. That convention was made to avoid randomized strategies in practice. For numerical purposes randomization gives a simplifying symmetry. Thus for this example when the point $(2, 2)$ is reached we have performed the computation as if we repaired the x -coordinate with probability .5 and repaired the y -coordinate with probability .5.

$$D = - \begin{bmatrix} 0 & 15 & 0 & 0 \\ 15 & 3 & 15 & 0 \\ 1 & 2 & 3 & 15 \\ 0 & 1 & 15 & 0 \end{bmatrix}, \quad T = \begin{bmatrix} 0 & 3 & 0 & 0 \\ 3 & 1 & 3 & 0 \\ 1 & 1 & 1 & 3 \\ 1 & 1 & 3 & 0 \end{bmatrix}.$$

The N matrix requires some computation. It is most easily determined by calculating the stationary probabilities for the corresponding two-dimensional random walk. Number the positive points in the following manner:

$$\begin{array}{ccccc} & & 9 & & \\ & 4 & 7 & 10 & \\ & 2 & 5 & 8 & 11. \\ & 1 & 3 & 6 & \end{array}$$

For this example the transition probability matrix $P = [p_{ij}]$ ($i, j = 1, \dots, 11$), where p_{ij} is the probability of a transition from the point numbered i to that numbered j , can be taken as

$$P = \begin{array}{c} \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{matrix} & \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix} \end{array}.$$

If $A = (a_1, a_2, \dots, a_{11})$ denotes the stationary probabilities, A can be determined from $AP = A$ and $\sum_{i=1}^{11} a_i = 1$. Because of the large number of 0's in P and the symmetry ($a_2 = a_3, a_4 = a_6, a_7 = a_8, a_9 = a_{11}$), the determination of A is not difficult. For this example,

$$A = \frac{1}{32} (6, 4, 4, 3, 4, 3, 2, 2, 1, 2, 1).$$

Then, since $n_i = a_i/a_1$,

$$N = \frac{1}{6} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 2 & 2 & 0 \\ 4 & 4 & 2 & 1 \\ 6 & 4 & 3 & 0 \end{bmatrix}$$

and

$$I = \frac{N \cdot D}{N \cdot T} = -3.423$$

Twenty-three strategies for the example, along with the corresponding N matrices and I 's, are given below. They are grouped according to $N(S)$, the number of elements in the C associated with S . Results of §§3, 4, 6 and 7 and the beginning of this section can be used to show that these 23 strategies form a complete class.

j	$N(S_j)$	S_j	N_j	$I(S_j)$
1	0	r_x	[1]	-5.000
2	1	r_y $c \quad r_x$	$\frac{1}{2} \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$	-3.750
3	3	r_y $c \quad r$ $c \quad c \quad r_x$	$\frac{1}{4} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 4 & 2 & 1 \end{bmatrix}$	-3.500
4	3	r_y $c \quad r_x$ $c \quad c \quad r_x$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 2 & 2 & 1 \end{bmatrix}$	-3.556
5	4	$r_y \quad r$ $c \quad c \quad r$ $c \quad c \quad r_x$	$\frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 4 & 2 & 1 \end{bmatrix}$	-3.385
6	4	$r_y \quad r_y$ $c \quad c \quad r_x$ $c \quad c \quad r_x$	$\frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}$	-3.400
7	5	r_y $c \quad r$ $c \quad r \quad r$ $c \quad c \quad c \quad r_x$	$\frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 4 & 1 & 0 \\ 8 & 4 & 2 & 1 \end{bmatrix}$	-3.500
8	5	r_y $c \quad r_y$ $c \quad r_x \quad r_x$ $c \quad c \quad c \quad r_x$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 4 & 1 & 0 \\ 2 & 4 & 2 & 1 \end{bmatrix}$	-3.957
9	6	r_y $c \quad r$ $c \quad c \quad r$ $c \quad c \quad c \quad r_x$	$\frac{1}{8} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 4 & 4 & 3 & 0 \\ 8 & 4 & 2 & 1 \end{bmatrix}$	-3.400
10	6	r_y $c \quad r_y$ $c \quad c \quad r_x$ $c \quad c \quad c \quad r_x$	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 4 & 4 & 3 & 0 \\ 2 & 4 & 2 & 1 \end{bmatrix}$	-3.429

\underline{j}	$\underline{N(S_j)}$	$\underline{S_j}$	$\underline{N_j}$	$\underline{I(S_j)}$
11	6	r		
(Optimal)		r_y c r	$\frac{1}{8} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 \\ 4 & 4 & 2 & 1 \\ 8 & 4 & 2 & 0 \end{bmatrix}$	-3.357
		c c c r		
		c c r_x		
12	6	r_y		
		r_y c r	$\frac{1}{6} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 \\ 4 & 4 & 2 & 1 \\ 6 & 4 & 2 & 0 \end{bmatrix}$	-3.360
		c c c r_x		
		c c r_x		
13	6	r_y		
		r_y c r_x	$\frac{1}{6} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 2 & 2 & 0 \\ 4 & 4 & 2 & 1 \\ 6 & 4 & 3 & 0 \end{bmatrix}$	-3.423
		c c c r_x		
		c c r_x		
14	7	r_y		
		c r	$\frac{1}{16} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 4 & 2 & 0 & 0 & 0 \\ 8 & 8 & 2 & 1 & 0 \\ 16 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.533
		c r		
		c r r r		
		c c c c r_x		
15	7	r_y		
		c r_y	$\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 4 & 2 & 0 & 0 & 0 \\ 8 & 8 & 2 & 1 & 0 \\ 2 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.641
		c r_y		
		c r_x r_x r_x		
		c c c c r_x		
16	8	r_y r		
		c c r	$\frac{1}{16} \begin{bmatrix} 2 & 3 & 0 & 0 \\ 4 & 6 & 6 & 0 \\ 8 & 8 & 6 & 3 \\ 16 & 8 & 4 & 2 \end{bmatrix}$	-3.364
		c c c r		
		c c c r_x		
17	8	r_y r_y		
		c c r	$\frac{1}{10} \begin{bmatrix} 2 & 3 & 0 & 0 \\ 4 & 6 & 6 & 0 \\ 8 & 8 & 6 & 3 \\ 10 & 8 & 4 & 2 \end{bmatrix}$	-3.368
		c c c r_x		
		c c c r_x		
18	8	r_y r_y		
		c c r_x	$\frac{1}{2} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 \\ 2 & 2 & 2 & 1 \\ 2 & 2 & 2 & 1 \end{bmatrix}$	-3.471
		c c c r_x		
		c c c r_x		
19	8	r_y		
		c r	$\frac{1}{16} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 4 & 6 & 0 & 0 & 0 \\ 8 & 8 & 6 & 1 & 0 \\ 16 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.438
		c r		
		c c r r		
		c c c c r_x		

j	$N(S_j)$	S_j	N_j	$I(S_j)$
20	8	r_y $c \quad r_y$ $c \quad r_y$ $c \quad c \quad r_x \quad r_x$ $c \quad c \quad c \quad c \quad r_x$	$\frac{1}{2}$ $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 4 & 6 & 0 & 0 & 0 \\ 8 & 8 & 6 & 1 & 0 \\ 2 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.488
21	10	r_y $c \quad r$ $c \quad c \quad r$ $c \quad c \quad c \quad r$ $c \quad c \quad c \quad c \quad r_x$	$\frac{1}{16}$ $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 \\ 4 & 6 & 6 & 0 & 0 \\ 8 & 8 & 6 & 4 & 0 \\ 16 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.400
22	10	r_y $c \quad r_y$ $c \quad c \quad r$ $c \quad c \quad c \quad r_x$ $c \quad c \quad c \quad c \quad r_x$	$\frac{1}{8}$ $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 & 0 \\ 4 & 6 & 6 & 0 & 0 \\ 8 & 8 & 6 & 4 & 0 \\ 8 & 8 & 4 & 2 & 1 \end{bmatrix}$	-3.414
23	10	r_y $c \quad r_y$ $c \quad c \quad r_x$ $c \quad c \quad c \quad r_x$ $c \quad c \quad c \quad c \quad r_x$	$\frac{1}{2}$ $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 4 & 4 & 4 & 3 & 0 \\ 2 & 4 & 4 & 2 & 1 \end{bmatrix}$	-3.543

The optimal strategy is number 11, for which $I^* = -3.357$ and C^* is not full. Several strategies are almost as good as the optimal strategy.

Glossary of symbols.

C_0 :	$\{z: i(z) \geq -mK/M\}$.
C_n :	net income from n th visited state.
$F_S(z)$:	$I_c(S z) - I^*T_c(S z)$.
$i(x, y)$ or $i(z)$:	income per unit of time while in the production state $z = (x, y)$ or the cost of repair for the repair state z .
$I(N S)$:	ratio of total expected income to total expected time from the first N visited states when using strategy S .
$I(S)$:	limit inferior of $I(N S)$.
I^* :	supremum of $I(S)$.
$I_c(z, S)$:	for the strategy S , the expected income until the first visit to the positive state z or until the cycle is over.
$I_c(S z)$:	for the strategy S , the expected income remaining in a cycle after first visiting the positive state z .
$I_c(S)$:	expected income during a cycle with the strategy S .

$K_0(K_x)[K_y]$:	cost per unit of time while repairing both $(x-)[y-]$ components.
$m_0(m_x)[m_y]$:	time required to repair both $(x-)[y-]$ components.
mK :	maximum (m_0K_0, m_xK_x, m_yK_y) .
M :	minimum $(m_0, m_x, m_y, 1)$.
$n_s(z)$:	for the strategy S , the expected number of visits per cycle to the state z .
p :	probability of a transition from (x, y) to $(x + 1, y)$ when control is not used.
$p_s(z)$:	for the strategy S , the steady state probability of the embedded Markov chain being at z .
$P_c(z, S)$:	for the strategy S , the probability that the state z will be visited during a cycle.
q :	probability of a transition from (x, y) to $(x, y + 1)$ when control is not used.
$r_0(r_x)[r_y]$:	the repair state for both $(x-)[y-]$ components. The same notation will represent z -points from which the corresponding repairs begin.
$R_0(R_x)[R_y]$:	the repair of both $(x-)[y-]$ components; also the set of z -points from which $R_0(R_x)[R_y]$ is begun.
S :	a strategy.
S^* :	an optimal strategy.
T_n :	time spent in n th visited state.
$T_c(z, S)$:	for the strategy S , the expected time until the positive state z is visited or until the end of the cycle.
$T_c(S z)$:	for the strategy S , the expected time remaining in a cycle after first reaching the positive state z .
$T_c(S)$:	expected length of a cycle with the strategy S .
$z = (x, y)$:	a possible production state when x and y are nonnegative integers or one of the repair states.
$z(t)$:	state of the process at time t .
z_n :	the n th visited state.

REFERENCES

- [1] CYRUS DERMAN, *On sequential decisions and Markov chains*, Management Sci., 9 (1962), pp. 16-24.
- [2] ———, Private communication, May 26, 1965.
- [3] I. RICHARD SAVAGE, *Surveillance problems*, Naval Res. Logist. Quart., 9 (1962), pp. 187-209.
- [4] V. S. TANEJA, *A surveillance model: two machine case*, Tech. Rep. 30, Department of Statistics, University of Minnesota, Minneapolis, 1963.

ON THE CONVERGENCE OF SOME FEASIBLE DIRECTION ALGORITHMS FOR NONLINEAR PROGRAMMING*

DONALD M. TOPKIS AND ARTHUR F. VEINOTT, JR.†

1. Introduction. Recently Goldstein [10], [11] has considered the problem of determining conditions on the selection of directions in feasible direction algorithms for unconstrained maximization problems which assure the gradient vanishes in the limit. This paper extends his work (when specialized to Euclidean spaces) in various ways, notably by considering constrained maximization problems and by allowing a wider choice in the selection of the step size.

We first give conditions on the directions and step sizes which assure convergence to a stationary point in feasible direction algorithms for maximizing a real valued continuous function on a closed set. Then we apply this result to establish convergence of a certain class of algorithms. In the unconstrained case the class includes Cauchy's [2] method of steepest ascents and the Newton-Raphson method, both with "optimal" (and certain other) step sizes. In the constrained case the class includes, among others, the method of Frank and Wolfe [7], Zoutendijk's procedure P2 [25], a simple variant of Zoutendijk's procedure P1 [25] which eliminates the need for his "anti-zigzagging" procedures, and some new second order methods which require quadratic programs to be solved at each stage. It is reasonable to conjecture that our second order methods for the constrained case will exhibit the same kind of computational superiority over first order methods such as those of Frank and Wolfe, Rosen [18], [19], and Zoutendijk, that the Newton-Raphson method for the unconstrained case exhibits over the method of steepest ascents. We are indebted to Robert Wilson for pointing out to us several years ago the importance of second order methods in nonlinear programming. The reader will notice that we have borrowed liberally from the terminology in Zoutendijk's important book [25].

2. Conditions assuring convergence to a stationary point. Let $F(\cdot)$ be a real valued continuous function defined on a subset S of E^N , N -dimensional Euclidean space. We seek $x \in S$, called *optimal*, that maximizes $F(\cdot)$ over S .

We say $d \in E^N$ is a *feasible direction* at $x \in S$ if for some $\delta > 0$ we have $x + sd \in S$ for all s , $0 < s \leq \delta$. If in addition $F(x) + s\delta < F(x + sd)$

* Received by the editors August 5, 1966, and in revised form December 2, 1966.

† School of Engineering, Program in Operations Research, Stanford University, Stanford, California. This research was supported by the Office of Naval Research under Contract Nonr-225(77).

for all s , $0 < s \leq \delta$, we say d is *usable* for $F(\cdot)$ at x .¹ If there is no usable direction for $F(\cdot)$ at $x \in S$, we say x is *stationary*. We remark that in many situations a stationary point will be optimal. For example, this is so if S is convex and $F(\cdot)$ is concave on S .

Let $x_0 \in S$ be given and let $S_0 = \{x: x \in S, F(x) \geq F(x_0)\}$. In the sequel we impose the following assumptions.

I. S is closed, $F(\cdot)$ is continuous on S , and S_0 is compact.

II. For every sequence $\{x_1, x_2, \dots\}$ in S_0 there is a bounded *direction function* $d(\cdot)$ which assigns to $\{x_0, x_1, \dots, x_n\}$ a feasible direction $d(x_0, x_1, \dots, x_n) = d_n$ at x_n , $n = 0, 1, \dots$. Moreover, there is a specified infinite set P of nonnegative integers such that any subsequence $\{(x_{n_k}, d_{n_k})\}$ of $\{(x_n, d_n): n \in P\}$ converging to (\bar{x}, \bar{d}) , say, has the property that

(a) for some $\delta > 0$, $x_{n_k} + sd_{n_k} \in S$ for $k = 1, 2, \dots$ and all s , $0 < s \leq \delta$, and

(b) if $\bar{x} \in S_0$ and \bar{d} is feasible but not usable at \bar{x} , then \bar{x} is stationary.²

III. There is a real valued lower semicontinuous *step size function* $f(\cdot, \cdot)$ defined on $S_0 \times S$ for which $f(x, x + \lambda d)$ is continuous in λ for fixed $x \in S_0$ and $d \in E^N$, and

(a) $f(x, y) \leq F(y)$ and $f(x, x) = F(x)$ for all $x \in S_0$, $y \in S$, and

(b) if d is a usable direction for $F(\cdot)$ at $x \in S_0$, then d is also usable for $f(x, \cdot)$ at x .

The algorithm. Under the hypotheses I–III we define x_1, x_2, \dots recursively by the rule

$$(1) \quad x_{n+1} = x_n + s_n d_n, \quad n = 0, 1, \dots,$$

where $d_n = d(x_0, x_1, \dots, x_n)$ and $s_n \geq 0$ is chosen so $x_n + s_n d_n \in S$. We require also that $F(x_{n+1}) \geq F(x_n)$ for $n \notin P$; and $f(x_n, x_n + s_n d_n) \geq f(x_n, x_n + sd_n)$ for all $s \geq 0$ with $x_n + sd_n \in S$ for $n \in P$.

As a simple illustration of the step size selection, if $f(x, y) = F(y)$ for $x, y \in S$ with $F(\cdot)$ continuous on S , then III holds. Also for $n \in P$, $s_n = s$ is then chosen to maximize $F(x_n + sd_n)$ subject to $s \geq 0$ and $x_n + sd_n \in S_0$. This selection of s_n is referred to as the *optimal step size*.

THEOREM 1. *If I–III hold, then the bounded sequence $\{x_n\}$ defined in (1) has the property that*

(i) $F(x_0) \leq F(x_1) \leq F(x_2) \leq \dots \leq \lim_{n \rightarrow \infty} F(x_n) < \infty$ and

(ii) *the limit of any convergent subsequence of $\{x_n: n \in P\}$ is stationary.*³

¹ If F is continuously differentiable on an open set containing S and d is feasible at $x \in S$, then d is usable for $F(\cdot)$ at $x \in S$ if and only if $\nabla F(x)d > 0$.

² Condition II(b) is a slightly weaker and reformulated version of a condition due to Goldstein [10], [11].

³ We remark that the limit points of $\{x_n: n \in P\}$ need not be stationary. However, all limit points of $\{x_n\}$ have the same function value.

Proof. From III for $n \in P$,

$$F(x_n) = f(x_n, x_n) \leq f(x_n, x_{n+1}) \leq F(x_{n+1}),$$

so (i) holds.

Let $\{(x_{n_k}, d_{n_k})\}$ be a convergent subsequence of $\{(x_n, d_n): n \in P\}$ with limit (\bar{x}, \bar{d}) . Since S is closed and there is a $\delta > 0$ such that II(a) holds, $\bar{x} + s\bar{d} \in S$ for $0 < s \leq \delta$ so \bar{d} is a feasible direction at \bar{x} .

From II(a) and III(a) we have, for $0 < s \leq \delta$,

$$\begin{aligned} f(\bar{x}, \bar{x}) = F(\bar{x}) &\geq \lim_{k \rightarrow \infty} F(x_{n_k+1}) \geq \lim_{k \rightarrow \infty} f(x_{n_k}, x_{n_k+1}) \\ &\geq \liminf_{k \rightarrow \infty} f(x_{n_k}, x_{n_k} + sd_{n_k}) \geq f(\bar{x}, \bar{x} + s\bar{d}). \end{aligned}$$

Thus from III(b), \bar{d} is not a usable direction for $F(\cdot)$ at \bar{x} . Consequently, by II(b), \bar{x} is stationary, which completes the proof.

The condition II(a) states that the directions $\{d_{n_k}\}$ are *uniformly* feasible at the points $\{x_{n_k}\}$. The condition is always satisfied if $S = E^N$ (as in unconstrained maximization problems) and so is then unnecessary. However, II(a) cannot be dispensed with entirely if $S \neq E^N$ as the following example shows.

Example in which II(a) fails to hold. Consider the problem of maximizing $F(y, z) \equiv z$ over $S \equiv \{(y, z): |y| \leq 1, |z| \leq 1, z + (1 - y^2)^{1/2} \geq 0\}$. The set S is the shaded area in Fig. 1. Let $x_{n-1} = ((1/n)(n^2 - 1)^{1/2}, -1/n)$, $d_n = \|x_{n+1} - x_n\|^{-1}(x_{n+1} - x_n)$, where $\|(y, z)\|^2 = y^2 + z^2$, and let s_n be chosen optimally. Then $\lim x_n = (1, 0) \equiv \bar{x}$ and $\lim d_n = (0, 1) \equiv \bar{d}$. Also, \bar{d} is usable at \bar{x} , so \bar{x} is not stationary. In this example all hypotheses of Theorem 1 are satisfied except for II(a).

The condition II(b) cannot be weakened to require only that d_{n_k} be usable for each k as we shall show in §3. Moreover, under mild additional hypotheses, II(b) *must* hold if the subsequences in II are to converge to stationary points from all starting points in S_0 . To see this suppose all hypotheses of Theorem 1 are satisfied except for II(b), \bar{d} (in II) not usable at \bar{x} implies $d(\bar{x}, \bar{x}, \dots, \bar{x})$ not usable at \bar{x} ,⁴ $s_n = 0$ whenever d_n is not usable, and \bar{d} is not usable at \bar{x} . Then starting at \bar{x} , the algorithm never leaves \bar{x} and so locates a stationary point only if \bar{x} is stationary.

In order to discuss applications of Theorem 1 it is convenient to introduce some notation. Let $C^k(X)$ be the set of real valued functions that are k times continuously differentiable on the open set $X \subset E^N$. Let $C^k(X)^m$ be the m -fold Cartesian product of $C^k(X)$. If $F \in C^1(X)$, let $\nabla F(x) = (F_{;i}(x))$ denote the gradient of F at x and let $\nabla^2 F(x) = (F_{ij}(x))$

⁴ This will be the case if, for example, $d(x_0, x_1, \dots, x_n) = d(x_n)$ and $d(\cdot)$ is continuous on S because then $\bar{d} = \lim d(x_{n_k}) = d(\bar{x})$.

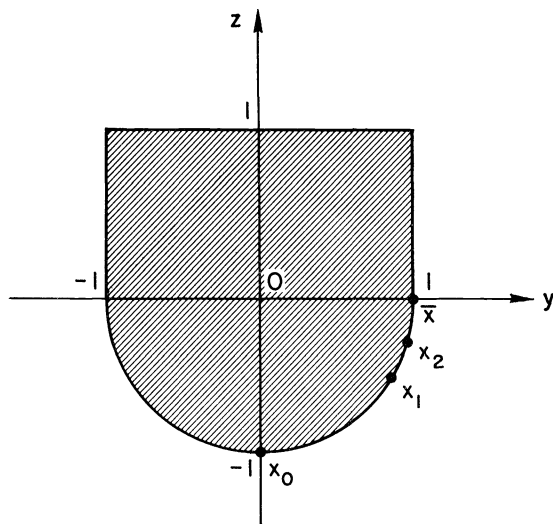


FIG. 1

denote the Hessian matrix of second partial derivatives of F at x . In the sequel $\nabla F(x)$ is a row vector and all other vectors are column vectors. If $G = (G^i) \in C^1(X)^m$, let $\nabla G(x)$ denote the matrix whose rows are $\nabla G^1(x), \dots, \nabla G^m(x)$.

3. Selection of the direction function. In this section we give some practical examples of direction functions satisfying II. We give results for the unconstrained case (Theorem 2) and for the constrained case (Theorem 3).

THEOREM 2. *If I and III hold, $S = E^N$, $F \in C^1(S)$, P is an infinite set of nonnegative integers, $\{H_n: n \in P\}$ is a bounded sequence of $N \times N$ matrices with every limit point of the sequence being negative definite, and $d_n = -H_n \nabla F(x_n)^T$ for $n \in P$, then each limit point \bar{x} of the bounded sequence $\{x_n: n \in P\}$ satisfies $\nabla F(\bar{x}) = 0$. Also, $\lim_{n \in P} \nabla F(x_n) = 0$.*

Proof. Since $\{(x_n, H_n): n \in P\}$ is bounded, there is a subsequence $\{(x_{n_k}, H_{n_k})\}$ such that $\lim_{k \rightarrow \infty} (x_{n_k}, H_{n_k}) = (\bar{x}, \bar{H})$. Now $F \in C^1(S)$ so $\lim_{k \rightarrow \infty} d_{n_k} = -\bar{H} \nabla F(\bar{x})^T \equiv \bar{d}$. Also \bar{d} is a feasible direction at \bar{x} and II(a) evidently holds. In addition, if \bar{d} is not usable at \bar{x} , then $0 \geq \nabla F(\bar{x}) \bar{d} = -\nabla F(\bar{x}) \bar{H} \nabla F(\bar{x})^T$, which implies that $\nabla F(\bar{x}) = 0$ because \bar{H} is negative definite. The theorem now follows from Theorem 1.

Examples and remarks. 1. *Cauchy's* [2] *method of steepest ascents.* If in Theorem 2 we let $P = \{0, 1, \dots\}$ and $H_n = -I$ for $n \in P$, then $d_n = \nabla F(x_n)^T$ for $n \in P$ and we have Cauchy's method.⁵

⁵ In the special case of Cauchy's method where s_n is chosen optimally, our proof reduces to one given by Ivanov [14].

2. *A modified Newton-Raphson method.* Suppose $F \in C^2(S)$ and $\nabla^2 F(x)$ is negative definite for $x \in S$. If in Theorem 2 we let $P = \{0, 1, \dots\}$ and $H_n = \nabla^2 F(x_n)^{-1}$ for $n \in P$, then $d_n = -\nabla^2 F(x_n)^{-1} \nabla F(x_n)^T$ for $n \in P$, and we have the Newton-Raphson method except for the manner of choosing the step sizes.

3. *An example with usable directions that fails to converge to a stationary point.* Consider the problem of choosing $(y, x) \in E^2$ that maximizes $F(y, z) = -\|(y, z)\|^2$ over $S \equiv E^2$ where $\|(y, z)\|^2 \equiv y^2 + z^2$. Let $O(r)$ be the circle centered at the origin and having radius $r \geq 0$. Let x_n be on the circle $O((n+2)/(n+1))$, $n = 0, 1, \dots$, where x_n is chosen for $n > 0$ so the line drawn from x_{n-1} through x_n is tangent to the circle $O((n+2)/(n+1))$ at x_n . Then $d_n = \|x_{n+1} - x_n\|^{-1} (x_{n+1} - x_n)$ is a usable direction at x_n , $n = 0, 1, \dots$, and the step size is optimal. Then the point $\bar{x} = \lim x_n$ lies on $O(1)$ so \bar{x} is not stationary or optimal. The difficulty is that each limit point \bar{d} of $\{d_n\}$ satisfies $\nabla F(\bar{x})\bar{d} = 0$ which illustrates the need for the hypothesis II(b) in Theorem 1 (all other hypotheses of the theorem are satisfied). This example shows, contrary to the claim of Householder [13, p. 134], that it does not suffice simply to pick any usable direction at each stage together with the optimal step size.

Since $\nabla F(x_n)d_n > 0$, $n = 0, 1, \dots$, in this example, it is easy to see that there exist negative definite symmetric matrices H_0, H_1, H_2, \dots such that $d_n = -H_n \nabla F(x_n)^T$, $n = 0, 1, \dots$. We may assume without loss of generality that $\{H_n\}$ is bounded (if not, normalize the $\{d_n\}$ and $\{H_n\}$, and adjust $\{s_n\}$ appropriately). However, it cannot occur that any limit point of $\{H_n\}$ is negative definite for we would then be assured by Theorem 2 that $\nabla F(\bar{x}) = 0$ which we have already seen does not occur. Thus the hypothesis in Theorem 2 that the limit points of $\{H_n\}$ are negative definite cannot be replaced by the weaker hypothesis that at most finitely many of the $\{H_n\}$ are not negative definite. In this connection see the remark on computational simplifications below.

4. *Acceleration devices.* According to Theorem 2, the directions d_n are specified only for $n \in P$. Thus for $n \notin P$, other choices of d_n which hopefully accelerate convergence are possible. For example, the steepest ascent partan method of Shah, Buehler and Kempthorne [21] is one possibility. Other possibilities are discussed in Spang [22, p. 350].

5. *Computational simplifications.* In order to simplify the Newton-Raphson method it is often desirable (as Robert Wilson has suggested to us) to compute the inverse of the Hessian matrix $\nabla^2 F(x_n)$ only occasionally. Then $H_n = \nabla^2 F(x_k)^{-1}$ for the largest $k \leq n$ for which the inverse is available. Under the remaining hypotheses of Example 2, Theorem 2 evidently still applies. Another simplification is to compute $\{H_n\}$ recursively in such a way as to "estimate" the desired inverse without ever computing

any second derivatives. A procedure for doing this is given in Fletcher and Powell [6]. However, in their method it remains an open question whether $\{H_n\}$ is bounded and has negative definite limit points. They establish convergence only for concave quadratic $F(\cdot)$.

In order to formulate algorithms for solving constrained maximization problems we impose the following assumption.

IV. Let $X, Y \subset E^N$ be open and $G = (G^i) \in C^1(Y)^m$. Assume $S = S' \cap S''$ where S', S'' are closed convex sets and $S'' \equiv \{x: G(x) \geq 0, x \in Y\}$ has an interior point a with $a \in S$. For each $i, 1 \leq i \leq m$, and $x \in S_0$ assume $G^i(x) = 0$ implies $\nabla G^i(x) \neq 0$. Also suppose $S \subset X$ and $F \in C^1(X)$.

We remark that if Y is convex and $G(\cdot)$ is quasi-concave on Y , then S'' will be convex. If in addition $G(\cdot)$ is concave on Y and for some $a \in Y, G^i(a) > 0$ for $1 \leq i \leq m$, then S'' has an interior point and $G^i(x) = 0$ implies $\nabla G^i(x) \neq 0$ for all i and $x \in Y$.

There is considerable freedom in the representation of a nonlinear program in the form IV. One attractive procedure is to partition the inequality constraints into two groups, linear and nonlinear. Then S' and S'' are respectively the sets of solutions to the linear and nonlinear inequality constraints. With this choice S' is a convex polyhedral set. We allow the possibility that $m = 0$ which means that $Y = S'' = E^N$.

THEOREM 3. *Suppose I, III, and IV hold, P is an infinite set of nonnegative integers, $\{H_n : n \in P\}$ is a bounded sequence of $N \times N$ matrices with every limit point of the sequence being negative semidefinite, $B \subset E^N$ is a compact convex neighborhood of the origin, and⁶*

$$g_n(x, y) = \min \{ \nabla F(x)(y - x) + \frac{1}{2}(y - x)^T H_n(y - x), [G(x) + \nabla G(x)(y - x)]^T \}$$

for $x \in S_0, y \in E^N$. Let y_n maximize $g_n(x_n, y)$ over $S' \cap (B + x_n)$, let θ_n be the largest number not exceeding one such that $x_n + \theta_n(y_n - x_n) \equiv z_n \in S''$, and let $d_n = z_n - x_n$ for $n \in P$. Then each limit point of the bounded sequence $\{x_n : n \in P\}$ is stationary.

Proof. Since $\{(x_n, H_n, y_n, \theta_n, z_n, d_n) : n \in P\}$ is bounded, there is a subsequence $\{(x_{n_k}, H_{n_k}, y_{n_k}, \theta_{n_k}, z_{n_k}, d_{n_k})\}$ converging to $(\bar{x}, \bar{H}, \bar{y}, \bar{\theta}, \bar{z}, \bar{d})$ with $\bar{x} \in S_0, \bar{y} \in S', \bar{z} \in S$. Now $x_{n_k}, z_{n_k} \in S$ and S is a closed convex set, so II(a) holds with $\delta = 1$ and \bar{d} is feasible at \bar{x} .

If \bar{d} is not usable at \bar{x} , then $\nabla F(\bar{x})\bar{d} \leq 0$. Also since \bar{H} is negative semidefinite, $\bar{d}^T \bar{H} \bar{d} \leq 0$. Thus upon letting $g(x, y) = \lim g_{n_k}(x, y)$, we see that $g(\bar{x}, \bar{z}) \leq 0$. Hence, by Lemma 4 below, $g(\bar{x}, \bar{y}) \leq 0$; so by Lemma 5 below,

⁶ $g_n(x, y)$ is the smallest component of the indicated $(m+1)$ -row vector. For $m = 0$, $g_n(x, y)$ is the first component in brackets.

\bar{x} is stationary.⁷ Therefore, II(b) holds and the theorem follows from Theorem 1.

LEMMA 4. *If $g(\bar{x}, \bar{z}) \leq 0$, then $g(\bar{x}, \bar{y}) \leq 0$.*

Proof. The proof is by contraposition. Suppose $g(\bar{x}, \bar{y}) > 0$. Thus $G^i(\bar{x}) = 0$ implies that

$$\nabla G^i(x)(y - x) \geq \epsilon$$

for some $\epsilon > 0$ for all (x, y) in some neighborhood of (\bar{x}, \bar{y}) . Hence for small enough η , $0 < \eta \leq 1$, and large enough K ,

$$G(x_{n_k} + s(y_{n_k} - x_{n_k})) \geq 0, \quad 0 < s \leq \eta, \quad k \geq K,$$

and so

$$x_{n_k} + s(y_{n_k} - x_{n_k}) \in S'', \quad 0 < s \leq \eta, \quad k \geq K.$$

It follows that $\bar{\theta} \geq \eta$, so, by the concavity of $g(\bar{x}, \cdot)$,

$$g(\bar{x}, \bar{z}) = g(\bar{x}, (1 - \bar{\theta})\bar{x} + \bar{\theta}\bar{y}) \geq (1 - \bar{\theta})g(\bar{x}, \bar{x}) + \bar{\theta}g(\bar{x}, \bar{y}) > 0,$$

which completes the proof.

LEMMA 5. *If $g(\bar{x}, \bar{y}) \leq 0$, then \bar{x} is stationary.*

Proof. The proof is by contraposition. Suppose \bar{x} is not stationary. Then there is $y \in S$ such that $\nabla F(\bar{x})(y - \bar{x}) > 0$. There is no loss in generality in assuming y is an interior point of S'' . For if not we may replace y by $y + \epsilon(a - y)$ for some small enough positive $\epsilon > 0$ and be assured that the desired inequality still holds, and that the new point is in S and in the interior of S'' . Now if $G^i(\bar{x}) = 0$, then $\nabla G^i(\bar{x})(y - \bar{x}) > 0$. For if not, since $\nabla G^i(\bar{x}) \neq 0$ by IV, there is a $z \in S''$ close enough to y such that $\nabla G^i(\bar{x})(z - \bar{x}) < 0$. But this means that $\bar{x} + \epsilon(z - \bar{x}) \notin S''$ for small enough positive ϵ , which contradicts the assumed convexity of S'' . Thus for each i , either

$$G^i(\bar{x}) > 0 \quad \text{or} \quad G^i(\bar{x}) = 0 \quad \text{and} \quad \nabla G^i(\bar{x})(y - \bar{x}) > 0.$$

Hence, for small enough positive ϵ , $g(\bar{x}, y') > 0$ where $y' = \bar{x} + \epsilon(y - \bar{x})$. It follows that

$$g(\bar{x}, \bar{y}) = \lim_{k \rightarrow \infty} g(x_{n_k}, y_{n_k}) = \max_{y \in S' \cap (B + \bar{x})} g(\bar{x}, y) > 0,$$

which completes the proof.

Examples and remarks. 1. *Reduction to Theorem 2.* Observe that if in

⁷ For $m = 0$, this assertion is easily proved as follows. We have $\bar{y} = \bar{z}$ so $g(\bar{x}, \bar{y}) \leq 0$ and \bar{d} is not usable. Also

$$\max_{y \in S' \cap (B + \bar{x})} g(\bar{x}, y) = \lim_{k \rightarrow \infty} g(x_{n_k}, y_{n_k}) = g(\bar{x}, \bar{y}) \leq 0,$$

so \bar{x} is stationary.

Theorem 3 we require that $m = 0$, $S' = E^N$, H_n be symmetric, the sequence $\{H_n^{-1}: n \in P\}$ be bounded with negative definite limit points, and B be large enough to contain the bounded sequence $\{-H_n^{-1}\nabla F(x_n)^T: n \in P\}$, then Theorem 3 reduces to Theorem 2 in the symmetric case. (Note H_n^{-1} in Theorem 3 becomes H_n in Theorem 2.)

2. *Method of Frank and Wolfe* [7] and *Zoutendijk's procedure* P2 [25]. If S' is a convex polyhedron, $B \supset (S' - S')$ (this implies $S' \cap (B + x) = S'$ for $x \in S'$), $P = \{0, 1, \dots\}$, and $H_0 = H_1 = \dots = 0$, the method of Theorem 3 becomes procedure P2 of Zoutendijk [25, p. 74]. If in addition $m = 0$, this procedure is the method of Frank and Wolfe [7].

3. *A variant of Zoutendijk's procedure* P1 [25]. If S' , B are convex polyhedrons, $P = \{0, 1, \dots\}$, and $H_0 = H_1 = \dots = 0$, the method of Theorem 3 becomes a variant of procedure P1 of Zoutendijk [25, p. 73]. His method may be described geometrically as follows. We may express S' as the intersection of finitely many closed half-spaces. The point x_n lies on the boundary of some (possibly none) of these half-spaces. Let S_n be the intersection of those particular half-spaces. Let $\tilde{g}_n(x_n, y)$ be the minimum of the numbers $\nabla F(x_n)(y - x_n)$ and $\{\nabla G^i(x_n)(y - x_n): G^i(x_n) = 0\}$. What Zoutendijk does is to choose $y = y_n$ to maximize $\tilde{g}_n(x_n, y)$ over $S_n \cap (B + x_n)$, i.e., he replaces g_n by \tilde{g}_n and S' by S_n . Zoutendijk was not able to prove that the limit points of $\{x_n\}$ are stationary under this procedure.⁸ Instead he complicates the stated method by introducing a rather unpleasant antizigzagging procedure which assures convergence. Our proposal (as given in Theorem 3) does not require any antizigzagging procedure.

The condition II(a) sheds some light on why Zoutendijk's method without the antizigzagging precaution fails. Suppose $\bar{x} \in S$ is on the boundary of S , the sequence $\{x_n\}$ is in the interior of S , and $x_n \rightarrow \bar{x}$. Also suppose B is the unit sphere, so $d_n = \nabla F(x_n)^T / |\nabla F(x_n)|^{-1}$ for all n . Then if \bar{x} is not stationary and $\nabla F(\bar{x})^T$ is not a feasible direction at \bar{x} , it is evident that II(a) cannot hold. It can be shown that the antizigzagging procedure assures that the directions $\{d_n\}$ are chosen such that II does hold.

4. *Solving the direction finding problem.* The direction finding problem of maximizing $g_n(x_n, \cdot)$ over $S' \cap (B + x_n)$ takes the form of a linear or quadratic program, or a sequence of quadratic programs when B is a convex polyhedron and S' is a convex polyhedral set.

If $H_n = 0$, the problem is equivalent to the linear program of choosing y, v that

$$(2) \quad \text{maximize } v$$

⁸ Indeed Zoutendijk [25, p. 72] states that examples in which the limit points are not stationary can be constructed. Recently Wolfe [24] has published such an example. Hadley [12, pp. 303-305] has given a fallacious convergence proof of Zoutendijk's procedure P1 for the case $m = 0$.

subject to

$$(3) \quad \begin{aligned} v &\leq \nabla F(x_n)(y - x_n), \\ v &\leq G^i(x_n) + \nabla G^i(x_n)(y - x_n), \quad 1 \leq i \leq m, \\ y &\in S' \cap (B + x_n). \end{aligned}$$

If B is "large", say $B \supset (S' - S')$, the last restriction in (3) simplifies to $y \in S'$. If instead B is "small", then an efficient computing procedure is to solve the (often trivial) problem of finding y that

$$(4) \quad \text{maximizes } \nabla F(x_n)(y - x_n) = v$$

subject to

$$(5) \quad y \in (B + x_n).$$

If y, v satisfy (3), the direction finding problem is solved. If not, the dual simplex method can be used effectively to solve (2), (3) starting with the solution to (4), (5). With this approach only the constraints in (3) that are "nearly active" at x_n will enter into the computations. Hence our method should require about the same computational effort as Zoutendijk's procedure P1. As a simple example, suppose $B = \{(w_j): -\eta \leq w_j \leq \eta, 1 \leq j \leq N\}$, where η is a small positive constant. Then one solution to (4), (5) is $y = x_n + r$, where $r = (r_j)$, $r_j = \pm\eta$ according as $\pm F_j(x_n) > 0$, and $r_j = 0$ if $F_j(x_n) = 0$. Other possible choices of B are discussed in Zoutendijk [25, p. 70].

If $m = 0$ and H_n is negative semidefinite, the direction finding problem is a concave quadratic program. If, instead, $m > 0$ and H_n is negative semidefinite, the direction finding problem may be solved by solving a sequence of quadratic programs of the following form. Choose y that

$$(6) \quad \text{maximizes } w = \nabla F(x_n)(y - x_n) + \frac{1}{2}(y - x_n)^T H_n (y - x_n)$$

subject to

$$(7) \quad \begin{aligned} v &\leq G^i(x_n) + \nabla G^i(x_n)(y - x_n), \quad 1 \leq i \leq m, \\ y &\in S' \cap (B + x_n). \end{aligned}$$

Let $w(v)$ denote the maximal value of w given v (≥ 0). If for some v (≥ 0) there is no y such that (v, y) satisfies (7), let $w(v) = -\infty$. Evidently $w(\cdot)$ is nonincreasing on $[0, \infty)$ and $w(0) \geq 0$. Therefore, $\varphi(v) \equiv w(v) - v$ is strictly decreasing in v and $w(w(0)) \leq w(0)$ so

$$\varphi(w(0)) \leq 0 \leq \varphi(0).$$

Thus, there is a unique number $v \in [0, w(0)]$ satisfying $\varphi(v+) \leq 0 \leq \varphi(v)$.

The interval bisection method may be used to find this number. Alternatively, a parametric quadratic programming method could be employed, for example, the long form of Wolfe's simplex method for quadratic programs applied to the dual of (6)–(7). Once it is found, the corresponding solution y to the quadratic program solves the direction finding problem.

5. *Some second order methods.* Suppose $F \in C^2(X)$, $\nabla^2 F(x)$ is negative semidefinite, and $H_n = \nabla^2 F(x_n)$ for $n \in P$. Then the algorithm of Theorem 3 is a second order method. It would appear to be especially attractive for $m = 0$. Since the modified Newton-Raphson method has been found superior to steepest ascents for unconstrained maximization problems, one may conjecture that our proposed second order methods will be superior to the first order methods discussed in Examples 2 and 3 above.

We remark that if $\nabla^2 G^i(\cdot)$ is negative semidefinite on Y for each i , then we may replace $G^i(x) + \nabla G^i(x)(y - x)$ by $G^i(x) + \nabla G^i(x)(y - x) + \frac{1}{2}(y - x)^T \nabla^2 G^i(x)(y - x)$ for each i in the definition of $g_n(x, y)$ in Theorem 3. No other changes are required in the statements or proofs of Theorem 3 or Lemmas 4 and 5. Although the quadratic approximation would probably speed up convergence of the algorithm, the direction finding problem then has quadratic constraints and hence is itself difficult to solve. One possibility is to solve the dual (in the sense of Stoer [23]) of this problem. The dual problem takes the form of minimizing a continuously differentiable convex function on the nonnegative orthant, provided $\nabla^2 F(\cdot)$ is negative definite on X and B, S' are polyhedral. Thus, algorithms like those discussed in Examples 2 and 3 above could be used to solve the dual of the direction finding problem.

6. Obvious analogues of the acceleration devices and computational simplifications discussed following Theorem 2 may be used here as well.

7. *Other methods.* There are a number of other methods which can be shown to satisfy the conditions of Theorem 1. One such class is the coordinate ascent methods. In these methods one only uses the coordinate directions. One possibility is to use each coordinate direction cyclically every N steps (D'Esopo [5], Ivanov [14] and Schecter [20]). Other possibilities are discussed in Spang [22, p. 345]. A second method is one devised by Zuhovickii, Polyak and Primak [26] for maximizing the minimum of finitely many continuously differentiable concave functions. A third method is Zoutendijk's procedure P1 [25]. (We remark that the second method seems to be equivalent to a special case of the third method.)

4. Selection of the step size function. In this section we give some practical examples of step size functions satisfying III. We have already discussed one example which led to the "optimal" step size.

1. *Quadratic approximation.* Suppose I holds, S is convex, and

$F \in C^2(E^N)$. Let $H(\cdot)$ be a function that assigns to each $x \in S$ a negative definite matrix $H(x)$. Assume $H(\cdot)$ is continuous on S . Let $\lambda(x, y) = \lambda$ be the largest number for which the determinant $|\nabla^2 F(y) - \lambda H(x)|$ vanishes, $x \in S_0$, $y \in S$. The characteristic value $-\lambda(x, y)$ has the property that

$$(8) \quad d^T[\nabla^2 F(y) - \theta H(x)]d \geq 0 \quad \text{for } \lambda(x, y) \leq \theta \quad \text{and } d \in E^N$$

by a theorem in Gantmacher [8, p. 319]. Thus if $\theta \geq \lambda(x, y)$ for all $x \in S_0$, $y \in S$, and $\theta > 0$, we may define the step size function as

$$(9) \quad f(x, y) = F(x) + \nabla F(x)(y - x) + \frac{\theta}{2} (y - x)^T H(x)(y - x).$$

It follows from (8) upon comparing (9) with the second order Taylor expansion of $F(y)$ about $F(x)$ that III(a) holds. Moreover, III(b) also holds and $f(\cdot, \cdot)$ is continuous on $S_0 \times S$, so III holds. Also $f(x, x + sd)$ is concave and quadratic in s and so is easy to maximize.

As a specific illustration of this technique, suppose $S = E^N$ and $H(x) = -I$. Then $-\theta$ is any negative number less than the smallest characteristic value of $\nabla^2 F(y)$ for each $y \in S$. For example, by a result of Barankin [1], θ may be any positive number as large as $\sum_j |F_{ij}(y)|$ for $i = 1, \dots, N$ and all $y \in S$. Also

$$s_n = \frac{1}{\theta} \frac{\nabla F(x_n) d_n}{d_n^T d_n}, \quad n = 0, 1, \dots$$

If, in addition, $d_n = \nabla F(x_n)^T$, then $s_n = 1/\theta$ (cf. Crockett and Chernoff [3]).

2. *Other methods.* There are a number of other ways of choosing step sizes that can be reduced to the principle of maximizing an appropriately chosen step size function satisfying III. Among these are the methods of Curry [4] and Goldstein [10], [11]. These particular methods lead to step size functions that are lower semicontinuous but not continuous.

3. *Approximately optimal step size.* Let $\{\eta_n : n \in P\}$ be a sequence of non-negative numbers such that $\lim \eta_n = 0$. Now suppose we modify the selection of the step size s_n in the algorithm (1) for $n \in P$ as follows: choose s_n so $F(x_{n+1}) \geq F(x_n)$ and $f(x_n, x_n + s_n d_n) \geq f(x_n, x_n + sd_n) - \eta_n$ for all $s \geq 0$ with $x_n + sd_n \in S$. Then Theorem 1 and its proof remain valid as is. One way of choosing s_n to achieve the above objective when $f(\cdot, \cdot)$ is continuous and I holds is so that $|s_n - s_n'| < \delta_n$, where $\delta_n \searrow 0$ and s_n' maximizes $f(x_n, x_n + sd_n)$ over $s \geq 0$, $x_n + sd_n \in S$.

REFERENCES

- [1] E. BARANKIN, *Bounds for the characteristic roots of a matrix*, Bull. Amer. Math. Soc., 51 (1945), pp. 767-70.

- [2] A. CAUCHY, *Méthode générale pour la résolution des systèmes d'équations simultanées*, C. R. Acad. Sci. Paris, 25 (1847), pp. 536-538.
- [3] J. CROCKETT AND H. CHERNOFF, *Gradient methods of minimization*, Pacific J. Math., 5 (1955), pp. 33-50.
- [4] H. CURRY, *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math., 2 (1944), pp. 258-261.
- [5] D. D'ESOP, *A convex programming procedure*, Naval Res. Logist. Quart., 6 (1959), pp. 33-42.
- [6] R. FLETCHER AND M. POWELL, *A rapidly convergent method for minimization*, Comput. J., 6 (1963), pp. 163-168.
- [7] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95-110.
- [8] F. GANTMACHER, *The Theory of Matrices*, vol. I, Chelsea, New York, 1959.
- [9] A. GOLDSTEIN, *Cauchy's method of minimization*, Numer. Math., 4 (1962), pp. 146-150.
- [10] ———, *On steepest descent*, this Journal, 3 (1965), pp. 147-151.
- [11] ———, *Minimizing functionals on normed linear spaces*, this Journal, 4 (1966), pp. 194-210.
- [12] G. HADLEY, *Nonlinear and Dynamic Programming*, Addison-Wesley, Reading, Massachusetts, 1964, pp. 296-305.
- [13] A. HOUSEHOLDER, *Principles of Numerical Analysis*, McGraw Hill, New York, 1953, pp. 132-134.
- [14] V. IVANOV, *A general approximation method for solving linear problems*, Soviet Math. Dokl., 3 (1962), pp. 415-418.
- [15] ———, *Algorithms of rapid descent*, Ibid., 3 (1962), pp. 476-479.
- [16] K. LEVENBERG, *A method for the solution of certain nonlinear problems in least squares*, Quart. Appl. Math., 2 (1944), pp. 164-168.
- [17] M. POWELL, *An efficient method for finding the minimum of a function of several variables without calculating derivatives*, Comput. J., 7 (1964), pp. 155-162.
- [18] J. ROSEN, *The gradient projection method for nonlinear programming. I. Linear constraints*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 181-217.
- [19] ———, *The gradient projection method for nonlinear programming. II. Non-linear constraints*, Ibid., 9 (1961), pp. 514-532.
- [20] S. SCHECTER, *Iteration methods for nonlinear programming*, Trans. Amer. Math. Soc., 104 (1962), pp. 179-189.
- [21] B. SHAH, R. BUEHLER AND O. KEMPTHORNE, *Some algorithms for minimizing a function of several variables*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 74-92.
- [22] H. SPANG, III, *A review of minimization techniques for nonlinear functions*, SIAM Rev., 4 (1962), pp. 343-365.
- [23] J. STOER, *Duality in nonlinear programming and the minimax theorem*, Numer. Math., 5 (1963), pp. 371-379.
- [24] P. WOLFE, *On the convergence of gradient methods under constraints*, IBM Research Report RZ 204, Zurich, Switzerland, 1966.
- [25] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960, Chap. 7.
- [26] S. ZUHOVICKII, R. POLJAK AND M. PRIMAK, *An algorithm for the solution of a problem of convex Chebyshev approximation*, Soviet Math. Dokl., 4 (1963), pp. 901-904.

THE MINIMIZATION OF A SMOOTH CONVEX FUNCTIONAL ON A CONVEX SET*

V. F. DEM'YANOV AND A. M. RUBINOV†

Abstract. A method is given for the minimization of a convex functional on a convex set in Banach space. The paper proves the convergence of the method and presents its applications to problems of optimal linear programming and of convex programming.

1. The minimization of a convex functional on a convex set.

1.1. Statement of the problem. We consider a Banach space X in which we are given a convex functional f having a gradient F . Let F have the derivative F' . (The definition of a gradient of a functional and of the derivative of an operator are given in [1].) Let Ω be a closed, bounded, convex set, $\Omega \subset X$. We are required to find $y \in \Omega$ such that

$$f(y) = \min_{x \in \Omega} f(x).$$

1.2. Certain additional information. Let $x, \tilde{x} \in \Omega$. We consider a functional f on the interval $[\tilde{x}, x]$, i.e., we consider a function $g(\alpha)$ defined for $\alpha \in [0, 1]$ in the following way:

$$(1.1) \quad g(\alpha) = f(\tilde{x} + \alpha(x - \tilde{x})).$$

From the properties of functional f we easily obtain that the function $g(\alpha)$ is twice continuously differentiable on $[0, 1]$ and, moreover,

$$(1.2) \quad \begin{aligned} g'(\alpha) &= (x - \tilde{x}, F(\tilde{x} + \alpha(x - \tilde{x}))), \\ g''(\alpha) &= (x - \tilde{x}, F'(\tilde{x} + \alpha(x - \tilde{x}))(x - \tilde{x})). \end{aligned}$$

(We denote the value of a functional $h \in X^*$ at an element $x \in X$ by (x, h) .)

Since f is convex, the function $(x - \tilde{x}, F(\tilde{x} + \alpha(x - \tilde{x}))) = g'(\alpha)$ does not decrease, while the function $(x - \tilde{x}, F'(\tilde{x} + \alpha(x - \tilde{x}))(x - \tilde{x})) = g''(\alpha)$ is not negative.

By a known formula in analysis,

$$\begin{aligned} g(\alpha) &= g(0) + \alpha g'(0) + o(\alpha); \\ g(\alpha) &= g(0) + \alpha g'(\theta), & \theta \in [0, \alpha]; \\ g(\alpha) &= g(0) + \alpha g'(0) + \frac{\alpha^2}{2} g''(\theta), & \theta \in [0, \alpha]. \end{aligned}$$

* Originally published in Vestnik Leningradskogo Universiteta, Seriya Matematiki, Mekhaniki i Astronomii, 19 (1964), pp. 5-17. Submitted on April 5, 1963. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Leningrad University.

If we replace the values of the function $g(\alpha)$ and of its derivatives by formulas (1.1) and (1.2), we have

$$(1.3) \quad f(\bar{x} + \alpha(x - \bar{x})) = f(\bar{x}) + \alpha(x - \bar{x}, F\bar{x}) + o(\alpha),$$

$$(1.4) \quad f(\bar{x} + \alpha(x - \bar{x})) = f(\bar{x}) + \alpha(x - \bar{x}, F(\bar{x} + \theta(x - \bar{x}))),$$

$$\theta \in [0, \alpha],$$

$$(1.5) \quad f(\bar{x} + \alpha(x - \bar{x})) = f(\bar{x}) + \alpha(x - \bar{x}, F\bar{x})$$

$$+ \frac{\alpha^2}{2}(x - \bar{x}, F'(\bar{x} + \theta(x - \bar{x}))(x - \bar{x})),$$

$$\theta \in [0, \alpha].$$

Formulas (1.3), (1.4) and (1.5) are called, respectively, the finite increment formula, the Lagrange formula and the Taylor formula.

1.3. Necessary and sufficient conditions for a minimum. Let $x \in \Omega$. We denote by \bar{x} any element at which the functional Fx attains a minimum on Ω . The following theorem is valid.

THEOREM 1. *In order for a convex, differentiable functional f to attain a minimum on Ω at the point y , it is necessary and sufficient that the point be a solution of the equation*

$$(1.6) \quad (\bar{x} - x, Fx) = 0.$$

Proof. Necessity. From the finite increment formula (1.3), for any $x \in \Omega$ and when $\bar{x} = y$, it follows that

$$(1.7) \quad f(y + \alpha(x - y)) - f(y) = \alpha(x - y, Fy) + o(\alpha).$$

Since the minimum is attained at the point y , the left-hand side of (1.7) is nonnegative when $\alpha \in [0, 1]$. For small α the sign of the right-hand side determined by the first term, whence $(x - y, Fy) \geq 0$. But when $x = y$ we have $(x - y, Fy) = 0$, and hence it follows that $\min_{x \in \Omega} (x - y, Fy) = (\bar{y} - y, Fy) = 0$.

Sufficiency. We assume that the assertion of the theorem is false. Let the point y satisfy (1.6), but let there exist an element $z \in \Omega$ such that $f(z) < f(y)$. From formula (1.9) when $\alpha = 1$, $x = z$, $\bar{x} = y$, we have

$$f(z) - f(y) = (z - y, F(y + \theta(z - y))), \quad \theta \in [0, 1],$$

whence we conclude that, since the function $(z - y, F(y + \alpha(z - y)))$ does not decrease,

$$0 > f(z) - f(y) \geq (z - y, Fy).$$

If we combine this inequality, $(z - y, Fy) < 0$, with the inequality $(\bar{y} - z, Fy) \leq 0$, which is valid by definition of element \bar{y} , we have

$(\bar{y} - y, Fy) < 0$. This contradicts the fact that y is a root of (1.6). The theorem is proved.

Remark. If the functional f is nonconvex, condition (1.6) is necessary but now not sufficient.

1.4. Method of successive approximations. It follows from Theorem 1 that to minimize the functional f on Ω it suffices to know how to solve (1.6). We cite here the method of successive approximations for the solution of this equation. The method is as follows.

1. As the first approximation we take an arbitrary point $x_1 \in \Omega$.

2. Suppose that the element x_n has already been constructed. We find \bar{x}_n and set up the function $g_n(\alpha)$, defined for $\alpha \in [0, 1]$, by means of (1.1),

$$g_n(\alpha) = f(\bar{x}_n + \alpha(x_n - \bar{x}_n));$$

$g_n(\alpha)$ is a convex function and reaches a minimum on $[0, 1]$ at some point which we shall denote by α_n .

Let us now set $x_{n+1} = \bar{x}_n + \alpha_n(x_n - \bar{x}_n)$. In this way we can construct the sequences

$$(1.8) \quad \begin{aligned} & x_1, x_2, \dots, x_n, \dots, \\ & \bar{x}_1, \bar{x}_2, \dots, \bar{x}_n, \dots. \end{aligned}$$

Here, as follows from the definition of x_{n+1} , we have

$$(1.9) \quad f(x_{n+1}) \leq f(x_n), \quad n = 1, 2, \dots$$

The following theorem is valid.

THEOREM 2. *Let f be bounded from below on Ω and let it have there a differentiable gradient F such that F' is bounded on Ω . Then*

$$\lim (\bar{x}_n - x_n, Fx_n) = 0.$$

Proof. Since f is bounded from below and since the sequence $f(x_n)$ is nonincreasing, the limit

$$\lim f(x_n) = Q > -\infty$$

exists.

Then, making use of the fact that $f(x_{n+1}) = \min_{\alpha \in [0,1]} f(\bar{x}_n + \alpha(x_n - \bar{x}_n))$ and of the Taylor formula, we have that for $\alpha \in [0, 1]$,

$$\begin{aligned} f(x_{n+1}) &\leq f(\bar{x}_n + \alpha(x_n - \bar{x}_n)) \\ &= f(x_n + (1 - \alpha)(\bar{x}_n - x_n)) \\ &= f(x_n) + (1 - \alpha)(\bar{x}_n - x_n, fx_n) \\ &\quad + \frac{1}{2}(1 - \alpha)^2(\bar{x}_n - x_n, F'(x_n + \theta(\bar{x}_n - x_n))(\bar{x}_n - x_n)), \\ &\hspace{15em} 0 \leq \theta \leq 1 - \alpha. \end{aligned}$$

Let $\sup_{x \in \Omega} \|F'x\| = M$. Then

$$\begin{aligned} & (\bar{x}_n - x_n, F'(x_n + \theta(\bar{x}_n - x_n))(\bar{x}_n - x_n)) \\ & \leq \|\bar{x}_n - x_n\| \cdot \|F'(x_n + \theta(\bar{x}_n - x_n))(\bar{x}_n - x_n)\| \\ & \leq \|F'(x_n + \theta(\bar{x}_n - x_n))\| \cdot \|\bar{x}_n - x_n\|^2 \\ & \leq MD^2, \end{aligned}$$

where $D = \sup_{z, z' \in \Omega} \|z - z'\|$ is the diameter of the set Ω .

Thus, for $\alpha \in [0, 1]$,

$$\begin{aligned} f(x_{n+1}) & \leq f(x_n) + (1 - \alpha)(\bar{x}_n - x_n, Fx_n) + \frac{1}{2}(1 - \alpha)^2 MD^2 \\ & = f(x_n) + (1 - \alpha)[(\bar{x}_n - x_n, Fx_n) + \frac{1}{2}(1 - \alpha)MD^2]. \end{aligned}$$

We note further that by the definition of the element \bar{x} we have

$$(\bar{x}_n - x_n, Fx_n) \leq 0, \quad n = 1, 2, \dots$$

We now suppose that the theorem is false. Then we can find a sequence of x_{n_k} and a $\rho > 0$, such that

$$(\bar{x}_{n_k} - x_{n_k}, Fx_{n_k}) \leq -\rho < 0, \quad k = 1, 2, \dots$$

In this case, when $\alpha \in [0, 1]$,

$$f(x_{n_k+1}) \leq f(x_{n_k}) + (1 - \alpha)[- \rho + \frac{1}{2}(1 - \alpha)MD^2].$$

Passing to the limit as $k \rightarrow \infty$, we have

$$Q \leq Q + (1 - \alpha)[- \rho + \frac{1}{2}(1 - \alpha)MD^2],$$

or, considering that $(1 - \alpha) > 0$,

$$- \rho + \frac{1}{2}(1 - \alpha)MD^2 \geq 0,$$

which is impossible when $0 \leq 1 - \alpha \leq 2\rho/MD^2$. The theorem is proved.

COROLLARY 1. *If Ω is compact and F is a differentiable operator with a derivative bounded on Ω , then the limit points of sequences (1.8) satisfy (1.6). If, here, f is a strictly convex functional, the sequences (1.8) converge.*

COROLLARY 2. *If Ω is weakly compact and F is a completely continuous and differentiable operator with a derivative bounded on Ω , then the limit points (in the weak topology) of sequences (1.8) satisfy (1.6). Strict convexity of f ensures the weak converges of sequences (1.8).*

THEOREM 3. *Let the convex functional f satisfy the hypotheses of Theorem 2 and let it attain a minimum on Ω at the point y . Then $f(x_n) \rightarrow f(y) = \min_{x \in \Omega} f(x)$.*

Proof. According to the Taylor formula (1.5), for any $x \in \Omega$ we have $f(x) - f(x_n) = (x - x_n, Fx_n) + (x - x_n, F'(x_n + \theta(x - x_n))(x - x_n))$.

Taking into account that $(x - x_n, F'(x_n + \theta(x - x_n))(x - x_n)) \geq 0$ we obtain

$$f(y) - f(x_n) = \min_{x \in \Omega} (f(x) - f(x_n)) \\ \geq \min_{x \in \Omega} (x - x_n, Fx_n) = (\bar{x}_n - x_n, Fx_n),$$

whence

$$(1.10) \quad 0 \leq f(x_n) - f(y) \leq (x_n - \bar{x}_n, Fx_n).$$

It now follows from Theorem 2 that $f(x_n) \rightarrow f(y)$; moreover, inequality (1.10) gives a convenient, a posteriori estimate of the convergence. The theorem is proved. We note that inequality (1.10) has been obtained by V. V. Khomenyuk from other considerations.

1.5. Modified method of successive approximations. When we apply the method presented in §1.4 for the minimization of a convex functional on a closed, bounded, convex set, we run into the known difficulties connected with the fact that to find the direction of descent accurately at each step, i.e., to find the element \bar{x}_n , and to determine the numbers α_n accurately (when minimizing the functions $g_n(\alpha) = f(\bar{x}_n + \alpha(x_n - \bar{x}_n))$), we may be required to carry out an infinite-step iteration process. We shall construct the successive approximations in the following manner.

Let $\min_{x \in \Omega} (x - x_n, Fx_n) = -\alpha_n$. In a finite number of steps we can determine the element \bar{x} satisfying only the condition $(\bar{x}_n - x_n, Fx_n) \in [-a_n, -a_n/2]$.

We now set

$$f(x_n) = b_n, \quad \min_{\alpha \in [0,1]} f(\bar{x}_n + \alpha(x_n - \bar{x}_n)) = c_n.$$

In a finite number of steps we can determine an $\alpha_n' \in [0, 1]$ such that $f(\bar{x}_n + \alpha_n'(x_n - \bar{x}_n)) \leq (b_n + c_n)/2$.

As x_{n+1} we take $x_{n+1} = \bar{x}_n + \alpha_n'(x_n - \bar{x}_n)$.

LEMMA 1. *Let all the hypotheses of Theorem 2 be fulfilled. Then*

$$\lim (\bar{x}_n - x_n, Fx_n) = 0.$$

Proof. We assume that the lemma is false. Then we can find a subsequence of n_k and a $\rho > 0$ such that

$$(\bar{x}_{n_k} - x_{n_k}, Fx_{n_k}) \leq -\rho < 0.$$

As in Theorem 2 it is easy to show that for any $\alpha \in [0, 1]$,

$$f(\bar{x}_n + \alpha(x_n - \bar{x}_n)) \leq f(x_n) + (1 - \alpha)[- \rho + \frac{1}{2}(1 - \alpha)MD^2].$$

We can find a $\rho_1 > 0$ such that when α is sufficiently close to unity, it turns out that $(1 - \alpha)[- \rho + \frac{1}{2}(1 - \alpha)MD^2] \leq -\rho_1 < 0$.

Let us further set $b_{n_k} = f(x_{n_k}) = Q + \epsilon_{n_k}$, where $\epsilon_{n_k} \rightarrow 0$, and hence, $\epsilon_{n_k} < \rho_1/2$ for k sufficiently large. For these n_k ,

$$f(\bar{x}_{n_k} + \alpha(x_{n_k} - \bar{x}_{n_k})) \leq Q + \epsilon_{n_k} - \rho_1 \leq Q - \frac{\rho_1}{2} < Q;$$

hence also

$$c_{n_k} = \min_{\alpha \in [0,1]} f(\bar{x}_{n_k} + \alpha(x_{n_k} - \bar{x}_{n_k})) \leq Q - \frac{\rho_1}{2}.$$

Thus, we have obtained

$$f(x_{n_k+1}) \leq \frac{b_{n_k} + c_{n_k}}{2} \leq \frac{Q + \epsilon_{n_k} + Q - \frac{\rho_1}{2}}{2} = Q - \frac{\rho_1}{4} + \frac{\epsilon_{n_k}}{2}.$$

Passing to the limit we have $Q < Q - (\rho_1/4)$, which is impossible since $\rho_1 > 0$. The following assertion is true.

ASSERTION. *Theorem 2 is valid for the modified process, i.e.,*

$$\lim (\bar{x}_n - x_n, Fx_n) = 0.$$

Proof. We assume the contrary and for some subsequence of x_{n_k} we let

$$\lim (\bar{x}_{n_k} - x_{n_k}, Fx_{n_k}) \leq -\rho < 0.$$

Then, for sufficiently large k we have

$$(x_{n_k} - x_{n_k}, Fx_{n_k}) = \min_{x \in \Omega} (x - x_{n_k}, Fx_{n_k}) \leq -\frac{\rho}{2} < 0.$$

Hence, if we take into account the rule for selecting the element \bar{x}_{n_k} , we have

$$(\bar{x}_{n_k} - x_{n_k}, Fx_{n_k}) \leq -\frac{\rho}{4} < 0,$$

which contradicts Lemma 1.

Remark. We can select \bar{x}_n from the condition $(\bar{x}_n - x_n, Fx_n) \in [-a_n, -a_n/s]$, $1 \leq s < \infty$. Analogously we can also select α'_n . Everything we have proved in §1.5 can be extended to this case.

2. The minimization of a convex functional on a linear set.

2.1. Statement of the problem. Let Ω be a closed, bounded, convex, balanced set in X . Let $S = L(\Omega)$ be the linear hull of Ω .

We are required to minimize a convex, differentiable functional f on S .

2.2. Necessary and sufficient conditions for a minimum.

THEOREM 4. *In order for f to attain a minimum on S at the point y , it is necessary and sufficient that*

$$(2.1) \quad (z, Fy) = 0$$

for any $z \in S$.

Proof. Necessity. From the finite increment formula we have

$$f(y + \alpha z) - f(y) = \alpha(z, Fy) + o(\alpha).$$

If the minimum is reached at the point y , the left-hand side is nonnegative. The sign of the right-hand side depends on the first term when α is small. However, since α is of arbitrary sign, $(z, Fy) = 0$.

Sufficiency. Let us assume that the theorem is false and that there exists a $z \in S$ such that $f(z) < f(y)$. Then, as was shown in the sufficiency proof of Theorem 1, $(z - y, Fy) < 0$. Since $(z - y) \in S$, the latter inequality contradicts (2.1).

THEOREM 5. *In order for f to attain a minimum on S at the point y , it is necessary and sufficient that this point be a root of the equation*

$$(2.2) \quad (\bar{x}, Fx) = 0,$$

where, as before, \bar{x} denotes the element of Ω at which $\min_{z \in \Omega} (z, Fx)$ is achieved.

Proof. Necessity. This follows from Theorem 4.

Sufficiency. Since the fact that the set Ω is balanced implies that together with the element x it also contains the element $-x$, from $(\bar{y}, Fy) = 0$ we have that

$$(x, Fy) \geq 0, \quad (-x, Fy) \geq 0,$$

which is possible only if $(x, Fy) = 0$. Since $S = L(\Omega)$, for any $z \in S$ we have $(z, Fy) = 0$. Sufficiency now follows from Theorem 4.

2.3. Method of successive approximations.

1. As the first approximation we take an arbitrary element $x_1 \in S$.

2. Suppose that the element x_n has already been chosen. We find \bar{x}_n and consider the element $x_{n\alpha} = x_n + \alpha \bar{x}_n$. We find the point $\alpha_n \in (-\infty, +\infty)$ at which $f(x_{n\alpha})$ attains a minimum. We now set

$$x_{n+1} = x_n + \alpha_n \bar{x}_n.$$

Thus, we have produced the sequences

$$(2.3) \quad x_1, x_2, \dots, x_n, \dots,$$

$$(2.4) \quad \bar{x}_1, \bar{x}_2, \dots, \bar{x}_n, \dots,$$

$$(2.5) \quad f(x_1), f(x_2), \dots, f(x_n), \dots.$$

Here, by construction,

$$f(x_1) \geq f(x_2) \geq \dots.$$

The method presented here is essentially a modification of the method of steepest descent proposed by L. V. Kantorovich [2]. In what follows we shall assume that

$$(2.6) \quad \inf_{z \in S} f(z) = Q > -\infty, \quad \sup_n \|x_n\| \leq R < +\infty.$$

The latter assumption is satisfied if, for example, f satisfies the following condition: for $z_n \in S$, $\sup_n \|z_n\| = \infty$ implies $\sup_n |f(z_n)| = \infty$.

THEOREM 6. *Let the functional f have a differentiable gradient F and, further, let F' be bounded on the sphere C_{R+D} of radius $R+D$, where R is determined from formula (2.6) and $D = \sup_{x \in \Omega} \|x\|$. Then $\lim (\bar{x}_n, Fx_n) = 0$.*

Proof. Using the Taylor formula, for any α we have

$$\begin{aligned} f(x_{n+1}) &= f(x_n + \alpha \bar{x}_n) \leq f(x_n + \alpha \bar{x}_n) \\ &= f(x_n) + \alpha(\bar{x}_n, Fx_n) + \frac{1}{2}\alpha^2(\bar{x}_n, F'(x_n + \theta \bar{x}_n)\bar{x}_n), \end{aligned}$$

where $0 \leq \theta \leq \alpha$.

In what follows we shall assume that $\alpha \in [0, 1]$ although $\alpha_n \in (-\infty, +\infty)$. Let

$$(2.7) \quad \sup_{x \in C_{R+D}} \|F'x\| = M.$$

Since $x_n + \theta \bar{x}_n \in C_{R+D}$ and $\bar{x}_n \in \Omega$, the inequality

$$(2.8) \quad f(x_{n+1}) \leq f(x_n) + \alpha(\bar{x}_n, Fx_n) + \frac{\alpha^2}{2} MD^2$$

is valid for $\alpha \in (0, 1)$.

Because Ω , together with any element z , also contains $-z$, it follows that $(\bar{x}_n, Fx_n) \leq 0$.

Let us assume that the theorem is false, and let a subsequence x_{n_k} and a $\rho > 0$ be found such that

$$(\bar{x}_{n_k}, Fx_{n_k}) \leq -\rho < 0.$$

Note further that, by assumption, $\lim f(x_n) = Q > -\infty$. Passing to the limit in (2.8) we have

$$Q \leq Q - \alpha\rho + \frac{1}{2}\alpha^2 MD^2 \quad \text{or} \quad -\rho + \frac{1}{2}\alpha MD^2 \geq 0,$$

which is impossible when $0 < \alpha < 2\rho/MD^2$.

COROLLARY. *If the sequences (2.3) and (2.4) have limit points, the limit points of sequence (2.3) are a solution of (2.2).*

Remark. As in §1.5 it is possible to propose a modified method of successive approximations.

3. Certain applications.

3.1. Solution of automatic control problems. Consider the linear system of n ordinary differential equations

$$(3.1) \quad \dot{x}(t) = A(t)x(t) + \sum_{i=1}^r B_i(t)u^i(t) + G(t),$$

with the initial conditions

$$(3.2) \quad x(0) = x_0.$$

Here $A(t)$ is an $n \times n$ matrix with real elements piecewise-continuous in $[0, T]$; $B_i(t)$, $i = 1, \dots, r$, and $G(t)$ are n -dimensional vectors with real components bounded on $[0, T]$; the r -dimensional control vector $u(t) = (u^1(t), \dots, u^r(t))$ is to be chosen from a certain bounded, convex, weakly-compact class U . For example, as U we can take one of the following classes: $u \in U$ if $u^i(t)$, $i = 1, \dots, r$, are bounded measurable functions given on $[0, T]$ and satisfying there one of the inequalities (3.3)–(3.6):

$$(3.3) \quad |u^i(t)| \leq 1, \quad i = 1, \dots, r, \quad t \in [0, T];$$

$$(3.4) \quad \int_0^T u^*(t)N(t)u(t)dt \leq 1,$$

where the asterisk denotes the transpose;

$$(3.5) \quad u^*(t)N(t)u(t) \leq \beta(t),$$

where $N(t)$ is an $r \times r$ matrix with real elements continuous on $[0, T]$ and is positive definite for every $t \in [0, T]$, and $\beta(t)$ is a real, nonnegative function piecewise-continuous on $[0, T]$;

$$(3.6) \quad \int_0^T u^{i2}(t)dt \leq \beta_i, \quad \beta_i > 0, \quad i = 1, \dots, r.$$

Let us choose any $u \in U$; the corresponding solution of system (3.1) with initial conditions (3.2), by the Cauchy formula, is

$$(3.7) \quad x(t, u) = Y(t)x_0 + \int_0^t Y(t)Y^{-1}(\tau) \left[\sum_{i=1}^r B_i(\tau)u^i(\tau) + G(\tau) \right] d\tau,$$

where $Y(t)$ is the transition matrix of system (3.1), $\dot{Y}(t) = A(t)Y(t)$, $Y(0) = E$, E is the $n \times n$ identity matrix. At each instant $t \in [0, T]$, (3.7) defines a set $R_t(U)$ in n -dimensional Euclidean space: $z \in R_t(U)$ if there exists $u \in U$ such that $x(t, u) = z$. We call this set the region of attainability for the system (3.1) at the instant t . From the properties of linear systems of differential equations, this set is convex for any t . Let the functional

$$(3.8) \quad J(u) = \int_0^T f(x(t, u), u(t), t)dt$$

be given, where $f(x, u, t)$ is a twice continuously differentiable function convex in x_i in the region of attainability of system (3.1) and convex in

u^j in the domain of u , continuous in t on $[0, T]$. We are required to find $\tilde{u} \in U$ such that

$$(3.9) \quad J(\tilde{u}) = \min_{u \in U} J(u).$$

We see from Theorem 1 that the necessary and sufficient conditions for a minimum have the following form in this case.

THEOREM 1'. *In order that the control $\tilde{u} \in U$ yield the minimum of functional (3.8), it is necessary and sufficient that*

$$(3.10) \quad \min_{u \in U} \int_0^T (u(t) - \tilde{u}(t))^* F(t, \tilde{u}) dt = 0,$$

where

$$F_i(t, \tilde{u}) = \omega^*(t, \tilde{u})[Y^{-1}(t)B_i(t)] + g_i(t, \tilde{u}),$$

$$\omega(t, \tilde{u}) = \int_t^T Y^*(\tau)C(\tau) d\tau,$$

$$C(\tau) = \left(\frac{\partial f(\tau, \tilde{u})}{\partial x_1(\tau, \tilde{u})}, \dots, \frac{\partial f(\tau, \tilde{u})}{\partial x_n(\tau, \tilde{u})} \right), \quad x = x_1, \dots, x_n,$$

$$g_i(t, \tilde{u}) = \frac{\partial f(t, \tilde{u})}{\partial u^i(t)}, \quad i = 1, \dots, r.$$

From (3.10) it is easy to obtain the maximum principle of L. S. Pontryagin [3] when the control class U satisfies inequality (3.3) or (3.5). The optimal control \tilde{u} can be found by the method of successive approximations presented in §1.4. In this case the controls \tilde{u}_n for minimizing the auxiliary linear functional are found at once in explicit form (for control classes U satisfying inequalities (3.3)–(3.6)). The method of successive approximations for the case of quadratic functionals has been discussed in [4]. We can find a necessary condition for a minimum, analogous to (3.10), for the case of nonlinear systems as well, and we can also propose a method of successive approximations similar to the one set forth in this paper; however, its convergence to a global minimum is not guaranteed in general because local minima may exist.

3.2. The solution of certain problems of approximation theory. Let X be a normed space in which we select a system of linearly independent elements

$$z_1, z_2, \dots, z_n.$$

Let Ω be the set of all polynomials of the form $P = \sum_{k=1}^n \alpha_k z_k$, where the coefficients α_k satisfy the following linear constraints:

$$\sum_{k=1}^n \alpha_{ik} \alpha_k \leq b_i, \quad i = 1, \dots, m.$$

Further, let $u \in X$. We are required to find $P_0 \in \Omega$ such that $f(u - P_0) = \min_{P \in \Omega} f(u - P)$, where f is a smooth, convex functional given on X . It is clear that to solve the stated problem it suffices to minimize (1.6) over this set.

We write this equation out explicitly for the case when

$$\Omega = \left\{ P = \sum_{k=1}^n \alpha_k z_k \mid |\alpha_k| \leq 1, k = 1, \dots, n \right\}.$$

Then, as is easily shown, for any

$$x = u - \sum_{k=1}^n \alpha_k z_k \in \Omega_u, \quad \tilde{x} = u - \sum_{k=1}^n \operatorname{sgn}(z_k, Fx) z_k.$$

Moreover,

$$\begin{aligned} (x - \tilde{x}, Fx) &= \left(u - \sum_{k=1}^n \alpha_k z_k - u + \sum_{k=1}^n \operatorname{sgn}(z_k, Fx) z_k, Fx \right) \\ &= \sum_{k=1}^n ((\operatorname{sgn}(z_k, Fx) - \alpha_k) z_k, Fx). \end{aligned}$$

As a result, (1.6) can be rewritten as

$$\sum_{k=1}^n ((\operatorname{sgn}(z_k, Fx) - \alpha_k) z_k, Fx) = 0.$$

We now consider the problem of best approximation under the assumption that there are absolutely no constraints imposed on the coefficients α_k . Again we denote

$$\Omega = \left\{ P = \sum_{k=1}^n \alpha_k z_k \mid |\alpha_k| \leq 1, k = 1, \dots, n \right\}.$$

Let $S = L(\Omega)$. In the present case our problem reduces to the minimization of the functional $f_u(x) = f(u - x)$ on the set S or, which is the same, to the solution of (2.2).

3.3. An integer programming problem. We are given a convex, twice-continuously-differentiable, real function $f(x)$ on a set H in the n -dimensional Euclidean space X . We are also given a finite point set L in H .

Problem A. Find a point $y \in L$ such that

$$(3.11) \quad f(y) = \min_{x \in L} f(x).$$

The search for points $y \in L$ satisfying (3.11) is a problem in integer programming.

To solve this problem we consider its connection to the noninteger (the so-called continuous) problem. We form the closed, bounded, convex set $\Omega \subset H: x \in \Omega$ if $x = \sum_{i=1}^s \alpha_i x_i$, $\sum_{i=1}^s \alpha_i = 1$, $\alpha_i \geq 0$, $x_i \in L$, $i = 1, \dots, s$.

Problem B. Find a point $y \in \Omega$ such that

$$f(y) = \min_{x \in \Omega} f(x).$$

For Problem B the necessary and sufficient conditions for a minimum take the form of the following theorem.

THEOREM 1''. *In order that the minimum of the function $f(x)$ on the set Ω be realized at a point $y \in \Omega$, it is necessary and sufficient that*

$$(3.12) \quad \min_{x \in \Omega} (x - y)^* F(y) = 0,$$

where $F(y) = (\partial f / \partial x_1, \dots, \partial f / \partial x_n) |_{x=y}$ is the gradient of function f at point y .

To find the points of minimum we can apply the method of successive approximations presented in §§1.4 and 1.5.

Let a point $y \in \Omega$ be known, for which $f(y) = \min_{x \in \Omega} f(x)$. To solve Problem A we use one of the following three methods.

1. Let there exist an algorithm for finding the integer points

$$(3.13) \quad x_{11}, \dots, x_{1p_1}, x_{21}, \dots, x_{2p_2}, \dots, \\ x_{ij} \in L, \quad i = 1, 2, \dots, \quad j = 1, \dots, p_i,$$

for which

$$(3.14) \quad (x_{11} - y)^* F(y) \\ = (x_{12} - y)^* F(y) = \dots = (x_{1p_1} - y)^* F(y) < \dots.$$

Then, by Theorem 1'',

$$(3.15) \quad (x_{ij} - y)^* F(y) \geq 0.$$

It may happen that $p_i = 1$ for some or for all i . Let

$$R_m = \{x_{11}, \dots, x_{1p_1}, x_{21}, \dots, x_{m1}, \dots, x_{mp_m}\},$$

$$\theta_m = \min_{z \in R_m} f(z),$$

$$\delta_m = (x_{mj} - y)^* F(y), \quad j = 1, \dots, p_m.$$

Then, for $z \in L$ but $z \notin R_m$,

$$(3.16) \quad (z - y)^* F(y) > \delta_m.$$

It is clear that as m increases, δ_m grows but θ_m perhaps only decreases. Let $\delta_m > 0$ (this is true when $m > 1$). Then

$$\begin{aligned}
 f(x) &= f(y + (x - y)) \\
 (3.17) \quad &= f(y) \\
 &\quad + (x - y)^* F(y) + (x - y)^* F'(y + \lambda(x - y))(x - y),
 \end{aligned}$$

$\lambda \in [0, 1]$, $2F'(\chi)$ is the matrix of second partial derivatives of the function f , computed at the point χ . For convex functions $f(x)$ the last term on the right-hand side of (3.17) is nonnegative, but then for $x \in L$, $x \notin R_m$, because of (3.16) it follows from (3.17) that

$$(3.18) \quad f(x) > f(y) + \delta_m.$$

If it turns out that $\theta_m \leq f(y) + \delta_m$, then $\theta_m = \min_{x \in L} f(x)$. If $f(x)$ is such that the matrix $F'(\chi)$ is strictly positive definite on Ω , i.e.,

$$(3.19) \quad z^* F'(\chi) z \geq r \|z\|^2, \quad \chi \in \Omega, \quad r > 0, \quad \|z\| = \sqrt{z^* z},$$

then for $x \in L$, $x \notin R_m$, from (3.16) we have

$$\delta_m < \|x - y\| \cdot \|F(y)\|, \quad \|x - y\| > \frac{\delta_m}{\|F(y)\|},$$

and for such x the inequality

$$f(x) > f(y) + \delta_m + \frac{r\delta_m^2}{\|F(y)\|^2}$$

follows from (3.17). Then, if

$$\theta_m \leq f(y) + \delta_m + \frac{r\delta_m^2}{\|F(y)\|^2},$$

then $\theta_m = \min_{x \in L} f(x)$.

2. If $f(x)$ is such that (3.19) is satisfied, and if we can obtain the integer points

$$(3.21) \quad x_{11}, \dots, x_{1p_1}, x_{21}, \dots, x_{2p_2}, \dots, \quad x_{ij} \in L,$$

such that

$$\begin{aligned}
 (3.22) \quad (x_{11} - y)^2 &= \dots = (x_{1p_1} - y)^2 \\
 &< (x_{21} - y)^2 = \dots = (x_{2p_2} - y)^2 < \dots,
 \end{aligned}$$

then we proceed as follows.

Let

$$Q_m = \{x_{11}, \dots, x_{1p_1}, \dots, x_{mp_m}\},$$

$$\theta_m = \min_{z \in Q_m} f(z),$$

$$\psi_m = (x_{mj} - y)^2, \quad j = 1, \dots, p_m.$$

It is clear that $\psi_m > 0$, $m = 1, 2, \dots$, and that for $x \in L$, $x \notin Q_m$: $(x - y)^2 > \psi$.

Since by Theorem 1" the second term on the right-hand side of (3.17) is nonnegative, for such x we have from (3.17) that

$$(3.23) \quad f(x) > f(y) + r\phi_m,$$

and if here $\theta_m \leq f(y) + r\phi_m$, then $\theta_m = \min_{x \in L} f(x)$.

3. If (3.19) is satisfied for $f(x)$ and if we can construct the sequences (3.13) and (3.21), then we can form the union $P_{ms} = R_m \cup Q_s$. Then, for $x \in L$, $x \notin P_{ms}$, we shall have

$$(3.24) \quad f(x) > f(y) + \delta_m + r\phi_s.$$

If $\theta_{ms} = \min_{x \in P_{ms}} f(x) \leq f(y) + \delta_m + r\phi_s$, then $\theta_{ms} = \min_{x \in L} f(x)$. It is not particularly difficult to construct the sequences (3.13) and (3.21) for a number of important practical problems. For example, consider the "traveling salesman problem". Given: $s + 2$ points in an n -dimensional space,

$$(3.25) \quad z_0, z_1, z_2, \dots, z_s, z_{s+1}.$$

Let us choose any set from these vectors,

$$(3.26) \quad \begin{aligned} & x_1, x_2, \dots, x_s, \\ & x_i \in \{z_1, \dots, z_s\}, \quad i = 1, \dots, s, \\ & x_j \in \{z_1, \dots, z_s\}, \quad j = 1, \dots, s. \end{aligned}$$

(there will be $s!$ such sets). We order these points,

$$z_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_s \rightarrow z_{s+1}.$$

This is a certain "path" in the n -dimensional space, whose length is

$$(3.27) \quad J(x_1, \dots, x_s) = \|z_0 - x_1\| + \sum_{i=1}^{s-1} \|x_{i+1} - x_i\| + \|z_{s+1} - x_s\|.$$

We are required to choose a set x_1, \dots, x_s such that the function (3.27) would take its least possible value.

The function $J(x_1, \dots, x_s)$ is not smooth and, therefore, in its place we consider a new function, which now is twice continuously differentiable and convex,

$$(3.28) \quad \begin{aligned} J_\epsilon(x_1, \dots, x_s) &= \sqrt{(x_1 - z_0)^2 + \epsilon} \\ &+ \sum_{i=1}^{s-1} \sqrt{(x_{i+1} - x_i)^2 + \epsilon} + \sqrt{(z_{s+1} - x_s)^2 + \epsilon}, \end{aligned}$$

where $\epsilon > 0$. For any $\delta > 0$ we can find an $\epsilon > 0$ such that $|J(x_1, \dots, x_s) - J_\epsilon(x_1, \dots, x_s)| \leq \delta$ for any set x_1, \dots, x_s .

At each step of the minimization by the method of successive approximations of (3.28) we must seek the minimum of the linear form

$$\sum_{i=1}^s x_i^* F_{im}, \quad x_i \in \{z_1, \dots, z_s\}, \quad i = 1, \dots, s,$$

$$z_j \in \{x_1, \dots, x_s\}, \quad j = 1, \dots, s,$$

and this is the allocation problem treated, for example, in [5].

REFERENCES

- [1] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.
- [2] L. V. KANTOROVICH, *Functional analysis and applied mathematics*, Uspehi Mat. Nauk, 3 (1948), pp. 89-185. (An English translation by C. D. Benster has appeared as NBS Rep. 1509, National Bureau of Standards, Los Angeles, California, 1952.)
- [3] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [4] V. F. DEM'YANOV, *The construction of an integral-optimal programmed control function in a linear system*, J. Appl. Math. Mech., 27 (1963), pp. 829-836.
- [5] D. B. YUDIN AND E. G. GOL'SHTEIN, *Problems and Methods in Linear Programming*, Fizmatgiz, Moscow, 1961.

TIME OPTIMAL CONTROL OF A LINEAR DIFFUSION PROCESS*

R. M. GOLDWYN†, K. P. SRIRAM‡ AND M. GRAHAM§

Abstract. The applicability of the Laplace transformation for the determination of the time optimal control of a linear diffusion process with amplitude constraints on the control is presented. The method—which can be interpreted as requiring a control whose transform in combination with the initial condition places zeroes at the poles of the open loop transfer function—is used to derive the optimal control function on the assumption that it is bang-bang, i.e., it is always at its limiting values. It is shown that the transfer of the system from a given initial state to a desired final state can be accomplished in finite time. A physical interpretation of the numerical results obtained is given, based on a transmission line analogue, and the actual time response for suboptimal controls is used to confirm theoretical estimates.

1. Introduction. In the past, much work has been done on the theory of the optimal control of lumped parameter systems, the motion of which may be described by means of ordinary differential equations. Many of the theoretical as well as the numerical results here are generally well known and one may refer to [9], for example, or the recent survey paper [12]. Equivalent results, particularly with specific examples, in the case of distributed parameter systems, the motion of which may be described by means of partial differential or integral equations, are not generally available. Some relevant papers on the control of distributed parameter systems are mentioned in [12]; for example, [10] gives a general view of the control problem, [8] gives a formulation of optimal processes in Banach spaces, and [5] specifically obtains numerical results by solving an integral equation. In this paper, the time optimal control of a one-dimensional heat diffusion process with amplitude constraints on the control is obtained making use of a transform technique. (As is known, there are certain advantages in using a transform such as the Laplace transform on linear distributed parameter systems. Representation of the Green's function is often easier in the transform domain than in the time domain and the transformed system re-

* Received by the editors August 17, 1966, and in revised form December 16, 1966. This work was supported in part by the National Science Foundation under Research Initiation Grant GP-3121 and under Grant GU-1153 with Rice University, Houston, Texas.

† Department of Electrical Engineering, Rice University, Houston, Texas. Now at the Thomas J. Watson Research Center, IBM, Yorktown Heights, New York.

‡ Department of Electrical Engineering, Rice University, Houston, Texas. Now at the Shell Development Company, Houston, Texas.

§ Department of Electrical Engineering, Rice University, Houston, Texas. Now at the Department of Electrical Engineering, University of California, Berkeley, California.

sponse is obtained as the solution of an ordinary differential equation with suitable boundary conditions.)

The most striking feature about the class of problems involving the transfer of a system from a given initial state to a desired stable final state, for convenience the origin, is that both the control and the state of the system are identically zero outside a finite time range. This fact, as will be shown later, dictates that the Laplace transform of both the control and the response be free of singularities in the finite complex s -plane. This requirement leads to the condition that the combination of the control and the initial condition must be such that they produce in the frequency domain zeroes at the poles of the open loop transfer function. The physical interpretation of this is interesting; the system has resonances at the poles of the open loop transfer function and the control is chosen so as to lead to a cancellation of these poles. In other words, the control forces the characteristic modes of the system to go to zero in finite time.

For the specific problem under consideration, two types of controls are considered: firstly, when the control is a temperature source, and secondly, when it is a heat flux source. On the assumption that the control is bang-bang, i.e., it is always at its limiting values, application of the condition that the poles be cancelled leads to an infinite set of equations for the switching times. A finite approximation to this infinite set is made and numerical results are obtained giving the control explicitly as a function of time. A physical interpretation of the results in terms of a transmission line analogue is presented. Because of the manner in which the control function is obtained, it is possible to make estimates of how close the actual response is to the desired response, and numerical solution for the response corresponding to different control functions confirms these estimates.

2. Finite time control and the Laplace transformation. The major results of this paper, in which the Laplace transformation is used on finite time control problems, are derived from the theorem given below.

THEOREM 1. *If $f(t)$ is a bounded, piecewise continuous function of t on $0 \leq t \leq T$ with a denumerable set of discontinuities and is identically zero for $t > T$, then the Laplace transform $\bar{f}(s)$ of $f(t)$,*

$$(1) \quad \bar{f}(s) = \int_0^{\infty} f(t)e^{-st} dt = \int_0^T f(t)e^{-st} dt,$$

is free of singularities in the entire finite complex s -plane. (Such a function is called an "entire" or "regular" function.)

It is easy to prove this theorem and, in fact, it follows from a more general result [6], according to which any function expressible as a finite integral whose integrand is a regular (analytic) function of s and continuous in the variable of integration is a regular function. It may be shown that the

result is also valid for piecewise continuous functions with a denumerable set of discontinuities. All that is necessary is to divide the range of integration into pieces, over each of which the integrand is a continuous function of the variable of integration. By the theorem in [6], each of the integrals thus obtained is an entire function, hence their sum is also entire. Necessary and sufficient conditions for $\bar{f}(s)$ ensuring that $f(t)$ is identically zero outside some finite time interval are given by Pfeiffer [11] and are quoted without proof in the Appendix.

The concept of entire functions may be directly applied to control problems such as those involving the transfer of a system from a given initial state to a desired final state (for convenience the origin) in finite time. The Laplace transform of the state of the system in such a finite time control problem must be entire since the state is identically zero after some finite time. Since the transform of the state is determined by the initial conditions, the control, and the differential equation governing the motion of the system, certain necessary conditions may be obtained which the control must satisfy, ensuring that the transform of the state is entire. In addition, since the control itself acts over a finite time, the transform of the control must also be entire.

3. Diffusion equation with temperature control. Consider the problem of one-dimensional heat diffusion in a uniform, homogeneous, isotropic body. The thickness of the body is normalized to unity and the end $x = 0$ of the body is insulated. At the end $x = 1$, the body is being heated (controlled) by a temperature source $f(t)$, where $f(t)$ is subject to the constraint $|f(t)| \leq 1$. At time $t = 0$, the body is at a uniform temperature of $+1$. The problem is to bring the body to a uniform temperature of zero in minimum time.

Denoting by $\theta(x, t)$ the normalized temperature at the point x at time t , the differential equation governing the system may be written as

$$(2) \quad \frac{\partial^2 \theta(x, t)}{\partial x^2} = \frac{\partial \theta(x, t)}{\partial t}.$$

In (2), time has been normalized to give a diffusivity of unity. The initial and boundary conditions may be represented mathematically as

$$(3) \quad \theta(x, 0) = 1,$$

$$(4) \quad \frac{\partial \theta}{\partial x}(0, t) = 0,$$

$$(5) \quad \theta(1, t) = f(t).$$

Taking the Laplace transform of (2) subject to (3) gives

$$(6) \quad \frac{\partial^2 \bar{\theta}(x, s)}{\partial x^2} = s\bar{\theta}(x, s) - 1,$$

where

$$(7) \quad \bar{\theta}(x, s) = \int_0^\infty \theta(x, t) e^{-st} dt.$$

The solution of (6) subject to (4) and (5) is

$$(8) \quad \bar{\theta}(x, s) = \frac{\cosh \sqrt{s} + \{s\bar{f}(s) - 1\} \cosh \sqrt{s}x}{s \cosh \sqrt{s}},$$

where

$$(9) \quad \bar{f}(s) = \int_0^\infty f(t) e^{-st} dt.$$

It is possible to specify certain conditions which the control must satisfy if the temperature is to be identically zero after a finite time. From the entire function concept, $\bar{\theta}(x, s)$ must be free of singularities in the finite complex s -plane. The denominator of (8) has an infinite set of zeroes at $s = s_n = -(n - \frac{1}{2})^2 \pi^2$. To have an entire function, the numerator must therefore be zero for these values of s . Hence the condition

$$(10) \quad s_n \bar{f}(s_n) - 1 = 0$$

must be satisfied.

4. Nature of optimum control function. Since the time optimal control for a linear finite state system is bang-bang, i.e., the control is always at its extreme values, the optimal control with $|f(t)| \leq 1$ is assumed to be of the form

$$(11) \quad f(t) = -u(t) + 2u(t - T_1) - 2u(t - T_2) + \cdots + (-1)^n u(t - T_n),$$

where $u(t)$ is the unit step function,

$$(12) \quad u(t) = \begin{cases} 0, & t < 0, \\ 1, & t \geq 0, \end{cases}$$

and T_1, T_2, \dots, T_n , are the times at which the control changes from -1 to $+1$ and vice versa. Obviously, the first switch would be negative with the initial condition as given. In the problem under consideration, there are an infinite number of singularities, so that, in (11), n would be infinite.

The Laplace transform of (11) then gives

$$(13) \quad \bar{f}(s) = -\frac{1}{s} [1 - 2e^{-sT_1} + 2e^{-sT_2} + \cdots].$$

Substitution of (10) in (13) gives an infinite number of equations for an infinite number of unknowns:

$$\begin{aligned}
 (14) \quad & 2 - 2e^{-s_1 T_1} + 2e^{-s_1 T_2} \dots = 0, \\
 & 2 - 2e^{-s_2 T_1} + 2e^{-s_2 T_2} \dots = 0, \\
 & \vdots \\
 & 2 - 2e^{-s_n T_1} + 2e^{-s_n T_2} \dots = 0, \\
 & \vdots
 \end{aligned}$$

It is required that $0 < T_1 < T_2 < \dots < T_n < \dots$. The s_i are all real and negative so that the left-hand side of any of equations (14) consists of an alternating series of increasing amplitude. The only way the series could have a finite sum is for successive terms to cancel each other.

Letting $T_{k+1} = T_k + \Delta T_k$, any two successive terms of the n th equation may be written as

$$(15) \quad e^{-s_n T_k} - e^{-s_n(T_k + \Delta T_k)} = e^{-s_n T_k} [1 - e^{-s_n \Delta T_k}],$$

so that convergence to a finite limit requires $-s_n \Delta T_k \rightarrow 0$ for large k and all n . Now, for large n , $|s_n| \rightarrow n^2$; hence it is necessary that $\Delta T_k \rightarrow 1/k^{2+\epsilon}$, $\epsilon \geq 0$ for large k , which implies that the series $\sum_{k=1}^{\infty} \Delta T_k$ converges to a finite limit.

It is easy to show that with the control of the form taken, $\bar{\theta}(x, s)$ satisfies the conditions given in the Appendix for being time limited when $0 \leq x \leq 1$.

5. Numerical solution of the equations. The exact system of equations (14) involves an infinity of switches, and direct numerical solution is impossible. However, finite approximations of the equation, involving truncation of the control after a finite number of switches can be solved, and then by increasing the number of switches, the solution to the exact system is approached. The finite approximation taking n switches is

$$\begin{aligned}
 (16) \quad & 2 - 2e^{-s_1 T_1} + 2e^{-s_1 T_2} - \dots + (-1)^n e^{-s_1 T_n} = 0, \\
 & 2 - 2e^{-s_2 T_1} + 2e^{-s_2 T_2} - \dots + (-1)^n e^{-s_2 T_n} = 0, \\
 & \vdots \\
 & 2 - 2e^{-s_n T_1} + 2e^{-s_n T_2} - \dots + (-1)^n e^{-s_n T_n} = 0.
 \end{aligned}$$

The control function for (16) is

$$f(t) = -u(t) + 2u(t - T_1) - 2u(t - T_2) + \dots - (-1)^n u(t - T_n),$$

and with n switches of this kind, (16) is the condition that the n poles closest to the origin are removed. Of course, it is possible to remove any n arbitrary poles, but the ones nearest the origin correspond to the most slowly decaying modes and it is desirable to remove these first. After n switches of this kind have been carried out, it can be asserted that whatever the remaining temperature might be, it would be decaying more rapidly than $e^{-(n+1/2)^2 \pi^2 t}$. Examination of (8) also reveals the fact that the residues at the poles go as

$$\frac{(-1)^n \cos(n - \frac{1}{2})\pi x}{(n - \frac{1}{2})\pi},$$

so that the magnitude of the remaining temperature would also be very small. Thus, even with a finite number of switches, it is possible to get arbitrarily close to the desired final state.

Numerical solutions of (16) for values of n between 1 and 18 are given in Table 1. The equations were solved using a double precision generalized Newton-Raphson method with scaling to take care of overflow problems on an IBM 7040 computer. The solutions obtained at each stage were used to get estimated starting values for the next stage.

It is interesting to note that the final switching times T_n appear to be converging. (The final switching time, T_n , for a particular n is the last value appearing for that n in Table 1.) This may be seen by multiplying the successive increments $\Delta T_n = T_n - T_{n-1}$ by n^2 , in which case the product $n^2 \Delta T_n$ decreases monotonically from 0.259771192 for $n = 2$ to 0.207321302 for $n = 18$. On this basis, a rough estimate of the final switching time as $n \rightarrow \infty$ may be made:

$$\begin{aligned} T_\infty &< T_{18} + 0.207321302 \sum_{i=19}^{\infty} \frac{1}{i^2} \\ (17) \quad &< T_{18} + 0.207321302 \int_{18}^{\infty} \frac{dx}{x^2} \\ &< 0.433525636. \end{aligned}$$

It should be noted that the $1/n^2$ variation is in accordance with what was indicated in the previous section.

When the initial temperature on the bar is not $+1$ but θ_0 , then a set of equations similar to (14) is obtained.

$$\begin{aligned} (18) \quad &1 + \theta_0 - 2e^{-s_1 T_1} + 2e^{-s_1 T_2} \dots = 0, \\ &1 + \theta_0 - 2e^{-s_2 T_1} + 2e^{-s_2 T_2} \dots = 0, \\ &\vdots \\ &1 + \theta_0 - 2e^{-s_n T_1} + 2e^{-s_n T_2} \dots = 0, \\ &1 + \theta_0 - 2e^{-s_{n+1} T_1} + 2e^{-s_{n+1} T_2} \dots = 0, \\ &\vdots \end{aligned}$$

θ_0 can be assumed to be positive without loss of generality. Once again, truncation of the switching function leads to the finite set of equations

$$(19) \quad 1 + \theta_0 + 2 \sum_{i=1}^{n-1} (-1)^i e^{-s_j T_i} + (-1)^n e^{-s_j T_n} = 0, \quad j = 1, 2, \dots, n.$$

TABLE 1
Switching times for temperature control, $\theta_0 = 1.0$

n	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	T_{11}	T_{12}	T_{13}	T_{14}	T_{15}	T_{16}	T_{17}	T_{18}
1	.280621																	
2	.314603	.345865																
3	.317902	.362194	.372340															
4	.318525	.365862	.382026	.386580														
5	.318698	.366260	.384765	.393028	.395460													
6	.318758	.366578	.385747	.395331	.400072	.401523												
7	.318782	.366709	.386157	.396301	.402006	.404991	.405926											
8	.318793	.366769	.386348	.396757	.402921	.406625	.408631	.409268										
9	.318799	.366799	.386445	.396902	.403393	.407470	.410023	.411438	.411892									
10	.318802	.366816	.386498	.397120	.403653	.407938	.410795	.412636	.413672	.414006								
11	.318804	.366825	.386529	.397194	.403804	.408211	.411249	.413339	.414712	.415493	.415747							
12	.318805	.366830	.386547	.397238	.403896	.408378	.411526	.413770	.415350	.416403	.417007	.417204						
13	.318806	.366834	.386558	.397266	.403954	.408483	.411702	.414046	.415758	.416984	.417890	.418286	.418443					
14	.318806	.366836	.386566	.397284	.403991	.408552	.411818	.414227	.416027	.417367	.418339	.418998	.419382	.419508				
15	.318806	.366838	.386571	.397296	.404017	.408599	.411895	.414349	.416210	.417627	.418698	.419482	.420018	.420331	.420433			
16	.318806	.366839	.386573	.397304	.404033	.408630	.411949	.414434	.416336	.417809	.418949	.419819	.420451	.420902	.421160	.421246		
17	.318806	.366840	.386576	.397310	.404045	.408652	.411987	.414494	.416426	.417937	.419127	.420059	.420776	.421309	.421677	.421893	.421964	
18	.318806	.366840	.386578	.397314	.404054	.408668	.412014	.414537	.416491	.418030	.419256	.420233	.420233	.421604	.422051	.422361	.422543	.422604

TABLE 2
Switching times for temperature control, $\theta_0 = 0.1$

n	T_1	T_2	T_3	T_4	T_5	T_6	T_7
1	.038628						
2	.065258	.090246					
3	.068498	.106125	.115701				
4	.069152	.109359	.125345	.129796			
5	.069336	.110287	.128113	.136240	.138644		
6	.069400	.110617	.129111	.138556	.143258	.144699	
7	.069427	.110753	.129529	.139534	.145198	.148169	.149100

Solutions of (19) for n up to 7 are given in Table 2. Proceeding as before, an estimate may be made of the final switching time as

$$(20) \quad \begin{aligned} T_\infty &< 0.149099511 + \frac{1}{7}(.215619296) \\ &< 0.179901981. \end{aligned}$$

Comparison of Tables 1 and 2 reveals the interesting fact that the spacing between the latter switches tends to become invariant irrespective of the initial temperature distribution. The initial temperature distribution affects only the first few switches appreciably, and conversely, the temperature distribution is affected most by the earlier switches.

6. Terminal time and time transmission line analogue. The diffusion equation considered above is a special case of the general transmission line equation

$$(21) \quad \frac{\partial^2 e}{\partial x^2} = LC \frac{\partial^2 e}{\partial t^2} + (LG + RC) \frac{\partial e}{\partial t} + RGe,$$

where R , L , C and G are the resistance, inductance, capacitance and conductance per unit length. Motivated by this transmission line analogue, it is reasonable to talk about the velocity of propagation of a temperature wave for the diffusion equation

$$(22) \quad \frac{\partial^2 \theta}{\partial x^2} = \frac{\partial \theta}{\partial t}.$$

If one examines (22) for a solution of the form

$$(23) \quad \theta = Ae^{i(kx + \omega t)}$$

with ω real, then such a solution represents a wave travelling in the negative x direction with a velocity $\omega/\text{Re}\{k\}$ and a decay factor of $\text{Im}\{k\}$. From (22) with (23),

$$(24) \quad -k^2 = i\omega$$

so that

$$(25) \quad \text{velocity of propagation} = \frac{\omega}{\operatorname{Re} \{k\}} = \sqrt{2\omega}.$$

The time taken, therefore, for a temperature wave of frequency ω to go from one end of the bar to the other is $\sqrt{1/(2\omega)}$ and the time for a round trip is $\sqrt{2/\omega}$. Hence given an arbitrary input function, the time for the entire slab to feel the complete effect of the control would be bounded by $\sqrt{2/\omega_s}$, where ω_s is the smallest frequency component of the control.

For $\theta_0 = 1$, the first switch occurs at about 0.3188 seconds, corresponding to the fundamental frequency of $\pi/0.3188$. A fair estimate of the final switch is then $\sqrt{(2 \times 0.3188)/\pi} = 0.45$ seconds. Similarly, for $\theta_0 = 0.1$, the final switch would occur at $\sqrt{(2 \times 0.0694)/\pi} = 0.21$ seconds. The estimates made in the previous section, viz., 0.433535636 and 0.179901981, are within these limits.

7. Diffusion equation with heat flux control. Consider next the same problem of one-dimensional heat diffusion with a different type of control: at the end $x = 1$ of the slab, instead of a temperature source, a heat source is applied. In terms of the transmission line analogue, this would correspond to driving the line with a current source rather than a voltage source.

The equations to be solved are now

$$(26) \quad \frac{\partial^2 \theta(x, t)}{\partial x^2} = \frac{\partial \theta(x, t)}{\partial x},$$

$$(27) \quad \theta(x, 0) = 1,$$

$$(28) \quad \frac{\partial \theta}{\partial x}(0, t) = 0,$$

$$(29) \quad \frac{\partial \theta}{\partial x}(1, t) = f(t).$$

Solving (26) through (29) for the transform of the temperature gives

$$(30) \quad \begin{aligned} \bar{\theta}(x, s) &= \frac{1}{s} + \frac{\bar{f}(s) \cosh \sqrt{s} x}{\sqrt{s} \sinh \sqrt{s}} \\ &= \frac{\sqrt{s} \sinh \sqrt{s} + s \bar{f}(s) \cosh \sqrt{s} x}{s \sqrt{s} \sinh \sqrt{s}}. \end{aligned}$$

At first glance, (30) appears to be of the same form as (8) and, in fact, the denominator of (30) has an infinite set of zeroes at $s_n = -n^2 \pi^2$. But there is one important difference in that the denominator here has a double zero at the origin. Hence the conditions to be satisfied are

$$(31) \quad s_n \bar{f}(s_n) = 0$$

and

$$(32) \quad \frac{d}{ds} [\sqrt{s} \sinh \sqrt{s} + s \bar{f}(s) \cosh \sqrt{s} x] \big|_{s=0} = 0,$$

to ensure that there is no double pole at $s = 0$. Assuming, as before, an optimum control with a finite number of switches of the form

$$(33) \quad f(t) = -u(t) + 2u(t - T_1) - \dots + (-1)^n u(t - T_n),$$

the transform of the control is

$$(34) \quad \bar{f}(s) = -\frac{1}{s} [1 - 2e^{-sT_1} + 2e^{-sT_2} - \dots + (-1)^n e^{-sT_n}].$$

Substitution of (34) in (32) gives

$$(35) \quad 1 - 2T_1 + 2T_2 - \dots + (-1)^n T_n = 0,$$

while substitution of (34) in (31) gives

$$(36) \quad \begin{aligned} 1 - 2e^{-s_1 T_1} + 2e^{-s_1 T_2} - \dots + (-1)^n e^{-s_1 T_n} &= 0, \\ \vdots \\ 1 - 2e^{-s_{n-1} T_1} + 2e^{-s_{n-1} T_2} - \dots + (-1)^n e^{-s_{n-1} T_n} &= 0. \end{aligned}$$

The physical interpretation of (35) is very interesting; it means that at whatever stage the truncation is carried out, the total heat input to the system must be -1 . Physically, this could have been anticipated since the net heat contained in the bar initially is $+1$ and at the end of the desired transition it is to be zero.

Numerical solutions of (35) and (36) are given in Table 3.

8. Time response with temperature control and flux control. The temperature in the slab as a function of x and t for a given temperature control may be obtained by taking the inverse Laplace transform of (8). Figure 1 gives the spatial distribution of temperature at different times corresponding to two switches. Figure 2 gives the time variation of temperature for a particular value of x , and here the effect of increasing the number of switches

TABLE 3
Switching times for flux control, $\theta_0 = 1.0$

n	T_1	T_2	T_3	T_4
1	1.000000			
2	1.070229	1.140458		
3	1.073595	1.164023	1.180858	
4	1.074076	1.167427	1.193317	1.199931

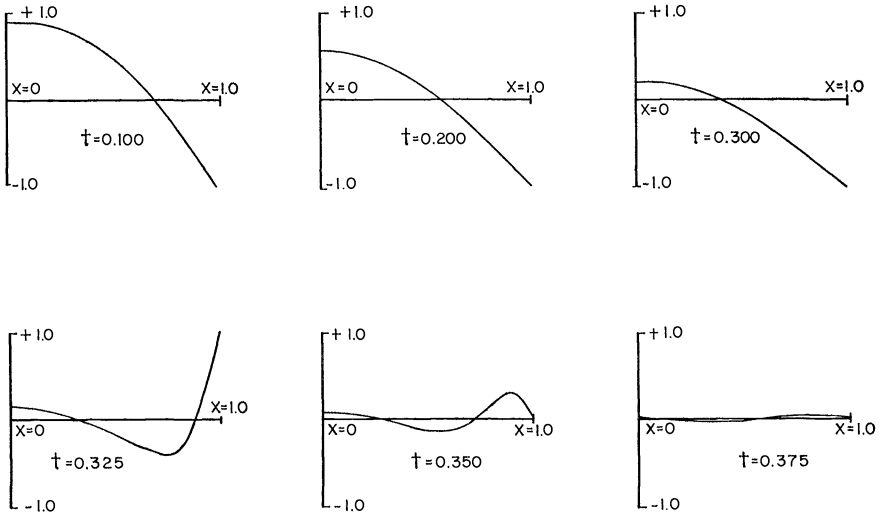


FIG. 1. Temperature control—spatial distribution of temperature. First switch at 0.31469278, second and final switch at 0.34586477.

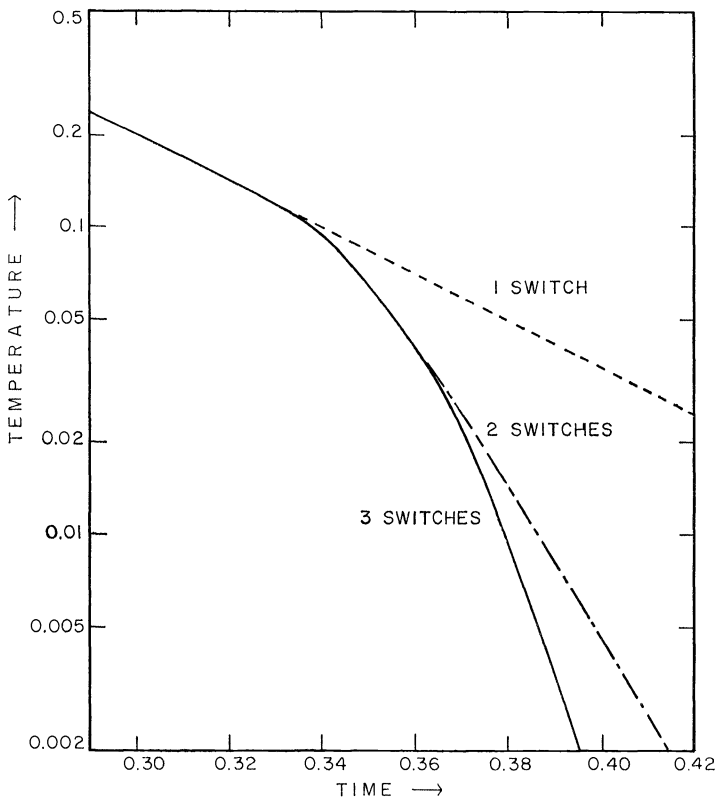


FIG. 2. Temperature control—time response at $x = 0.1$

is clear. The final decay after the switching is stopped is exponential in every case, and by increasing the number of switches more of the modes nearer to the origin are cancelled. For instance, for $n = 1$, the time constant of the decay after the final switch is close to $4/(9\pi^2)$, while for $n = 2$, it is $4/(25\pi^2)$, and so on for n switches. This was anticipated in §5. In this way it is possible to get arbitrarily close to the origin with a finite number of switches.

Figures 3 and 4 give similar plots of temperature when the heat flux at the end of the bar is being controlled. The sharp discontinuities of tempera-

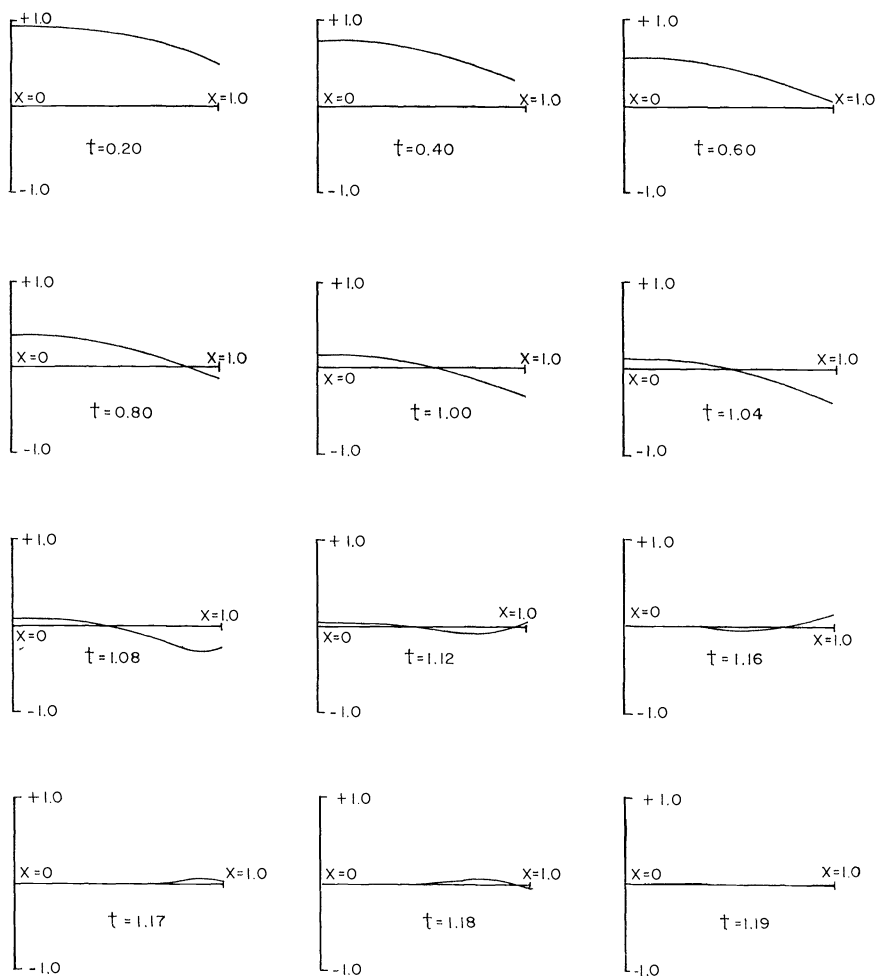


FIG. 3. Flux control—spatial distribution of temperature. First switch at 1.0735946, second switch at 1.1640234, third and final switch at 1.1808575.

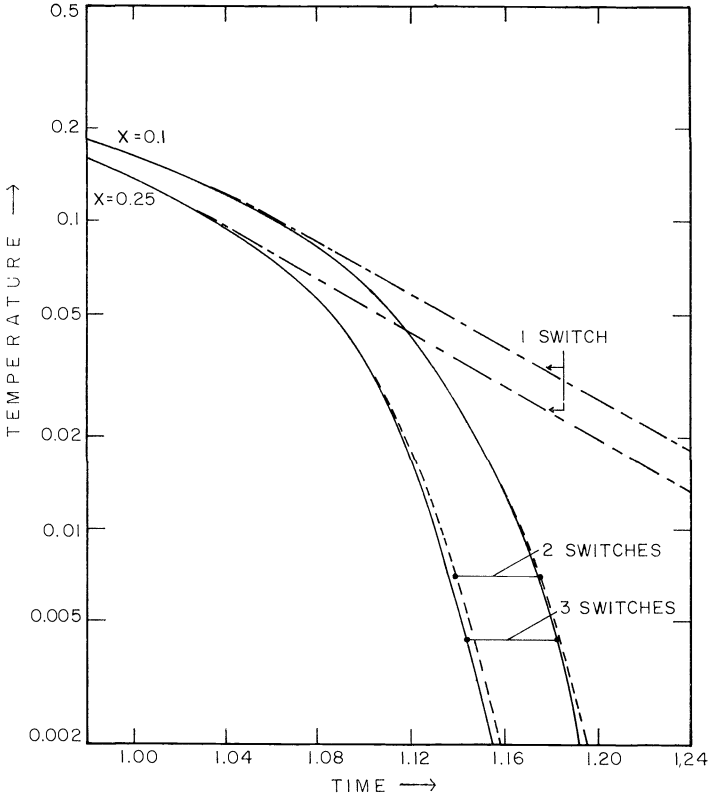


FIG. 4. Flux control—time response at $x = 0.1$ and $x = 0.25$

ture at the instants of switching are absent, but the effect of increasing the number of switches on the final decay is still evident.

Conclusion. The applicability of the Laplace transformation to determine the time optimal control of a linear diffusion process has been demonstrated. Work on the general transmission line equation and examination of the optimal control of hyperbolic systems is being undertaken. Another promising prospect is the application of transform techniques to system parameter identification when the representation of the Green's function—as is often the case—is easier in the transform domain than in the time domain. A time integral may be transformed by Parseval's theorem into an integral over all frequencies.

Appendix.

THEOREM [11]. A necessary and sufficient condition that $f(t)$, an arbitrary, bounded continuous function of t on $0 \leq t \leq T$, be identically zero for $t > T$

is that the Laplace transformation $\bar{f}(s)$ of $f(t)$,

$$\bar{f}(s) = \int_0^\infty f(t)e^{-st} dt = \int_0^T f(t)e^{-st} dt,$$

satisfy the following conditions:

- (i) $\bar{f}(s)$ has no singularities in the finite complex s -plane;
- (ii) there is a positive M such that

$$|\bar{f}(s)| < Me^{|\sigma|T}, \text{ where } \sigma = \operatorname{Re} \{s\};$$

- (iii) $|\bar{f}(s)| \rightarrow 0$ as $s \rightarrow \infty$ along any path in $\operatorname{Re} \{s\} \geq \sigma_0$ for any $\sigma_0 > -\infty$.

REFERENCES

- [1] YU. V. EGOROV, *Certain problems in the theory of optimal control*, Soviet Math. Dokl., (1962), pp. 1080-1084.
- [2] A. G. BUTKOVSKII AND A. YA. LERNER, *Über die optimale Steuerung von Systemen mit verteilten Parametern*, Regelungstechnik, 1 (1961), pp. 185-188.
- [3] I. McCausland, *Time optimal control of a linear diffusion process*, Proc. Inst. Elec. Engrs. (Science and General), 112 (1965), pp. 543-548.
- [4] A. G. BUTKOVSKII, *The method of moments in the theory of optimal control of systems with distributed parameters*, Automat. Remote Control, 24 (1963), pp. 1106-1113.
- [5] Y. SAKAWA, *Solution of an optimum control problem in a distributed parameter system*, IEEE Trans. Automatic Control, AC-9 (1964), pp. 420-426.
- [6] E. T. COPSON, *Functions of a Complex Variable*, Oxford University Press, London, 1960, p. 107.
- [7] A. I. EGOROV, *Optimal processes in systems containing distributed parameter plants (I and II)*, Automat. Remote Control, 26 (1965), pp. 972-988, 1178-1187.
- [8] A. V. BALAKRISHNAN, *Optimal control in Banach spaces*, this Journal, 3 (1965), pp. 152-180.
- [9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [10] P. K. C. WANG AND F. TUNG, *Optimal control of distributed parameter systems*, Proceedings of the Joint Automatic Control Conference, University of Michigan, 1963, pp. 16-32.
- [11] P. E. PFEIFFER, *The Laplace transform of functions which vanish outside an interval*, Series EE 66, Rep. 20, Rice University, Houston, Texas, 1966.
- [12] M. ATHANS, *The status of optimal control theory and applications for deterministic systems*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 580-596.

ON THE PROBLEM OF DEDUCING STATES AND STATE-RELATIONS FROM INPUT-OUTPUT RELATIONS FOR LINEAR TIME-VARYING SYSTEMS*

A. V. BALAKRISHNAN†

1. Introduction. This paper deals with the state-space theory of continuous systems as developed by Zadeh [1] and the author [2]-[4]. One of the central problems in this theory is that of determining states from the input-output relations. (For the definition of system and state and related concepts used herein, see [1].) It is not known at present whether it is always possible to do this. In this paper we introduce a sufficient condition on the input-output relations and show how from this property it is possible to deduce state-spaces and associated relations for a class of time-varying systems.

We call the property “weak null connectedness” or “weak null controllability.” When specialized to a linear dynamic system, this is recognized as related to the well-known condition [1], [5] for complete controllability or complete algebraic connectedness. By a dynamic system we mean here that the input $u(t)$ and the output $y(t)$ satisfy equations of the form:

$$\begin{aligned}\dot{x}(t) &= A(t)x(t) + B(t) u(t), \\ y(t) &= C(t) x(t) + \sum_0^m d_k(t)u^{(k)}(t); \end{aligned}$$

and conversely, any pair of functions $u(t)$, $y(t)$ that satisfy these equations are admissible input-output pairs, it being assumed that the first order ordinary differential equation has a unique solution for each initial value. As is known (see [1], for example), these may then be taken as state-input and state-input-output relations. Our condition for weak null connectedness is given without any reference to states and is based only on the input-output relation. The state-spaces that we deduce cannot be and are not restricted to be finite-dimensional. This is true also of the reduced state-spaces; they may not be even normable, although we show that there is a natural topology for them which makes them locally convex. Conditions are given for the reduced state-spaces to be finite-dimensional. The Zadeh result [1], [2] that a linear time-invariant system with a finite-dimensional state-space is dynamic (or can be represented as a dynamic system) is shown to be incorrect in the time-varying case. It is shown to be true for

* Received by the editors May 16, 1966, and in revised form March 6, 1967.

† Department of Engineering, University of California, Los Angeles, California 90024. This research was supported in part by the Air Force Office of Scientific Research, United States Air Force, under Grant 700-67.

a weakly null connected system whose reduced state-spaces are all of the same finite dimension.

For recent related results for time-varying systems (with finite-dimensional state-spaces) with potential applications to lumped-parameter circuit theory, see [8] and [9], with additional references given in the latter.

We begin with a definition of weak null connectedness in terms of states for a linear dynamic system and discuss the relationship to complete controllability. It turns out that the two are equivalent in the time-invariant case, but neither necessarily implies the other in the time-varying case.

2. Definition of weak null connectedness. Let us first consider a linear "dynamic" system, where the input $u(t)$ (assumed to be one-dimensional for simplicity of notation, the extension to the general finite-dimensional case being obvious) and the output $y(t)$ (also assumed one-dimensional) are related by the equations:

$$(1) \quad \frac{dX(t)}{dt} = A(t)X(t) + B(t)u(t),$$

$$(2) \quad y(t) = C(t)X(t) + \sum_0^m d_k(t)u^{(k)}(t),$$

where $X(t)$ is an $n \times 1$ matrix function, $A(t)$ is an $n \times n$ matrix function, $B(t)$ is $n \times 1$, $C(t)$ is $1 \times n$, and $d_k(t)$ are scalar, all functions measurable and essentially bounded on finite intervals. For such a system, the state-space at any instant of time can be taken to be (isomorphic to) an n -dimensional vector space, denoted E_n . In fact, since the assumptions on $A(t)$ and $B(t)$ imply that (1) has a unique solution for each initial value (as a Cauchy problem), we may take (1) as the state-input relation and (2) as the state-input-output relation. With reference to (1), we may state our definition for weak null connectedness or weak null controllability (abbreviated WNC) as follows.

DEFINITION. The system defined through (1) and (2) is said to be *weakly null connected* or *weakly null controllable* if, given any state X at any time t , it is possible to find an earlier instant of time t_0 , $t_0 < t$, and an input starting at time t_0 , such that (1) has a solution satisfying

$$X(t_0) = 0, \quad X(t) = X.$$

To see why this definition is new and different, let us recall the usual definition of *complete controllability* or *complete connectedness* [1], [5] (abbreviated CC): Given any state X_1 at time t_1 and any other state X_2 , it is possible to find an input which will transfer the state X_1 at time t_1 to the state X_2 at some $t > t_1$; that is, we can find a solution of (1) satisfying

$$X(t_1) = X_1 \quad \text{and} \quad X(t) = X_2 \quad \text{for some} \quad t > t_1.$$

THEOREM 2.1. *Suppose the system (1) is time-invariant, i.e.,*

$$(3) \quad dA(t)/dt = 0, \quad dB(t)/dt = 0.$$

Then weak null controllability is equivalent to complete controllability.

Proof. Since (3) holds, let

$$(4) \quad A(t) = A, \quad B(t) = B.$$

The proof is immediate since we have then that

$$\exp At = \sum_0^{n-1} Q_k(t) A^k$$

and

$$\int_{t_0}^t (\exp A(t-s)) Bu(s) ds = \sum_0^{n-1} \left(\int_{t_0}^t Q_k(t-s) u(s) ds \right) A^k B,$$

so that weak null connectedness readily implies that $A^k B$ be linearly independent for $k = 0, \dots, n-1$. The latter being the well-known condition for complete controllability, the proof is clearly complete.

However, Theorem 2.1 is false in general in the time-varying case, as the following simple counterexample shows. Thus, let (1) have the special form

$$\frac{dX}{dt} = AX(t) + B(t)u(t),$$

where

$$(5) \quad B(t) = \begin{cases} 0 & \text{for } t < \text{some fixed } t_0, \\ B & \text{for } t > t_0, \end{cases}$$

and $\{A^k B\}$ are linearly independent, $k = 0, \dots, n-1$. Then for any τ , fixed and less than t_0 , we have

$$\int_{\tau}^t (\exp A(t-s)) B(s) u(s) ds = \int_{t_0}^t \exp A(t-s) Bu(s) ds,$$

and by taking t larger than t_0 , we see that any state can be reached, so that we have CC. On the other hand, no nonzero state can be reached at any time t , $t < t_0$, starting with the zero state at any earlier time, or we do not have WNC. By reversing the inequality in (5) it is clear that we can obtain a system that is WNC, but not CC.

However, a sufficient condition, essentially a sort of analyticity condition on $A(t)$ and $B(t)$ that has played a key role in optimal control theory and was invoked by Pontryagin, Boltyanskii, Gamkrelidze and Mishchenko [6], may be given implying CC and WNC simultaneously. (This in [6] is

called the "general position condition.") Thus, let P denote the following condition.

The vectors $B_k(t)$ are linearly independent for every t :

$$(6) \quad \begin{aligned} B_k(t) &= -A(t)B_{k-1}(t) + dB_{k-1}(t)/dt, \quad k = 1, \dots, n-1, \\ B_0(t) &= B(t), \end{aligned}$$

it being assumed, of course, that the necessary differentiability conditions are satisfied.

It may also be noted that if a system (1) is WNC, then for any t , any state at time t can be reached from the zero state starting at the fixed time $t - L$, for some fixed positive L ; this is simply a consequence of the finite dimensionality of the state-space. We can also give a set of necessary and sufficient conditions for a dynamic system (with the state-input relation (1)) to be WNC. For this let $Q_i(s)$ be the column vectors of the matrix $Y(s)$ which is a fundamental matrix solution of the equation

$$\frac{dY(s)}{ds} = -A^*(s)Y(s).$$

Then we can state the following theorem.

THEOREM 2.2. *A necessary and sufficient condition for the system (1) to be WNC is that*

$$(7) \quad \sum_1^n a_i B(s)^* Q_i(s) = 0, \quad s < t \text{ for some (any) } t,$$

implies that all a_i are zero.

Proof. Suppose the system is WNC. Then it is not possible to find a nonzero vector Y , such that

$$(8) \quad Y^* \int_{-\infty}^t \psi(t)^{* -1} \psi(s) B(s) u(s) ds = 0$$

for every $u(\cdot)$, or equivalently that

$$(9) \quad B(s)^* \psi(s) \psi(t)^{-1} Y = 0, \quad s < t,$$

and since $\psi(t)$ is nonsingular, this implies the condition (7). Conversely, if (7) holds, we obtain equivalently (8) and hence (9), which would then yield WNC.

Remark. It may be noted that no differentiability condition has been used in (7), but the fundamental matrix solution is required. We can, of course, state a corresponding (one-sided) version of P in place of (7) with appropriate differentiability conditions. Also, we obtain CC by reversing the inequality in (7).

3. Definition of WNC based on input-output relation. Weak null controllability (as well as complete controllability) has been defined in terms of the state-input relation for a linear dynamic system. Such a definition can be clearly extended to any system with a state-space description even including those whose state-spaces are not finite-dimensional, but it requires the notion of state. We shall now give a definition of WNC which is based only on the input-output relation, and without reference to states. In particular, therefore, such a definition can be applied to systems which are described only in terms of the input-output relation. We shall show that for systems satisfying the new definition of WNC, it is possible to deduce states and associated state-space description from the input-output relation, and further that the states then will have the WNC property also.

Our definition of weak null connectedness is essentially that the output is uniquely specified for $t > t_0$ by specifying a previous input history prior to t_0 . There is, in other words, a previous input history, which may be unknown but which, if given, will determine the output corresponding to the input for all the future. More specifically, we define a system given in terms of input-output relations to be *linear and weakly null controllable* if the output $y(t)$, corresponding to the input $u(t)$ for $t > t_0$, can be expressed in the form:

$$(10) \quad y(t) = \int_{t_0-L}^{t_0} W(t, s)u_0(s) \, ds + \int_{t_0}^t W(t, s)u(s) \, ds,$$

$$0 < L \leq \infty,$$

for some $u_0(\cdot)$, where $W(t, s)$ is a fixed function of the two variables s, t , $-\infty < s, t < +\infty$. The interpretation of the integral as well as the precise definition of $L = +\infty$ depends on the kind of restrictions placed on the function $W(\cdot, \cdot)$. If we allow $W(t, s)$ to be a generalized function, then the input functions have to be infinitely differentiable, and the output functions to be generalized functions also. On the other hand, the class of input functions can be extended to integrable functions if we restrict $W(s, t)$ to be, say, Lebesgue measurable and essentially bounded on finite intervals. We shall assume this in what follows. In particular then, the integral is defined almost everywhere in t for finite L for inputs integrable on finite intervals, and the limiting case of $L = \infty$ will be interpreted to mean that the integral converges almost uniformly in each interval of the form $[t_0, t_0 + \Delta]$, $0 < \Delta < \infty$.

If a linear dynamic system is WNC, then it is clearly also WNC in the new definition. For, from (1), we have that

$$x(t) = \Phi(t)\Phi(s)^{-1}x(t_0) + \int_{t_0}^t \Phi(t)\Phi(s)^{-1}B(s)u(s) \, ds,$$

where $\Phi(t)$ is the fundamental matrix solution of

$$\Phi(t) = A(t)\Phi(t);$$

and if the system is WNC, we must have

$$x(t_0) = \int_{t_0-L}^{t_0} \Phi(t_0)\Phi(s)^{-1}B(s)u_0(s) \, ds$$

for some (finite) $L > 0$ and appropriate $u_0(\cdot)$. Hence, from (2),

$$y(t) = \int_{t_0-L}^{t_0} W(t, s)u_0(s) \, ds + \int_{t_0}^t W(t, s)u(s) \, ds,$$

where

$$W(t, s) = C(t)\Phi(t)\Phi(s)^{-1}B(s) + \sum_0^m d_k(t)\delta^k(t-s),$$

and $\delta^k(\cdot)$ denotes the k th derivative of the delta function. The first term is measurable and essentially bounded on finite intervals, while the second term requires that the generality of Schwartz distributions be allowed. By restricting ourselves to the former class we forego the inclusion of the derivatives of the input functions.

We shall now consider the problem of determining states and state-relations for a system whose input-output relation satisfies the WNC criterion (10). An obvious implication of (10) is, of course, that the output is determined if we can specify the input history, and hence the latter should qualify as states. The states are no longer finite-dimensional, and the precise definitions and, more important, the precise relations are given in the following theorem.

THEOREM 3.1. *Let the input-output relation be specified by (10), where we assume that $W(t, s)$ is Lebesgue measurable and essentially bounded on finite intervals. Then it is possible to deduce state-spaces for the system, such that the reduced states are WNC. The reduced state-spaces can be taken as locally convex spaces, with the state-input relation being given by*

$$x(t) = \int_{t_0}^t T(t; s)B(s)u(s) \, ds + T(t; t_0)x(t_0),$$

where $T(t; s)$ is a family of linear continuous transformations, such that

$$T(t; \tau)T(\tau; s) = T(t; s), \quad s < \tau < t,$$

$B(s)$ is a reduced state for each s , and the output-state relation is $y(t) = C(t)[x(t)]$ a.e., where $C(t)$ is a linear functional on the reduced state-space at time t .

Proof. Because of our assumptions on the weight function $W(t, s)$, the

class of inputs for which the output is defined as a measurable function includes the linear class of functions of bounded variation (or set functions countably additive on bounded Borel sets) with compact support in $(-\infty, +\infty)$. We shall denote this class by \mathfrak{B} . For the input β in \mathfrak{B} the output is given by

$$(11) \quad y(t) = \int_{-\infty}^t W(t, s) d\beta(s),$$

defined almost everywhere in t , $-\infty < t < +\infty$. Since we will be dealing for the most part with absolutely continuous (set) functions and with (set) functions with purely atomic parts, we shall henceforth write (11) more simply as

$$(12) \quad y(t) = \int_{-\infty}^t W(t, s) u(s) ds,$$

allowing for delta functions in $u(\cdot)$ as necessary. To obtain a state-space representation we begin by rewriting (10) as

$$(13) \quad \begin{aligned} y(t) &= \int_{-\infty}^{t_0} W(t, s) u(s) ds + \int_{t_0}^t W(t, s) u(s) ds \\ &= \int_{-\infty}^{t_0} W(t_0 + \Delta, t_0 + s) u(t_0 + s) ds + \int_{t_0}^{t_0 + \Delta} W(t, s) u(s) ds, \\ &\quad t_0 < t = t_0 + \Delta. \end{aligned}$$

Let us denote by Σ the subclass of functions in \mathfrak{B} whose support is confined to $(-\infty, 0]$. We note that the "past input" with reference to (13), $u(t_0 + s)$, $-\infty < s \leq 0$, is an element of Σ . Let us now introduce a family of linear transformations $L(t)$, $-\infty < t < +\infty$, defined on Σ , mapping Σ into \mathfrak{L} , the space of Lebesgue measurable functions on $(0, \infty)$ which are essentially bounded on finite intervals, by means of

$$L(t)u = v, \quad v(\Delta) = \int_{-\infty}^0 W(t + \Delta, t + s) u(s) ds, \quad \Delta \geq 0.$$

Then, of course, $L(t)$ is linear. We shall show that Σ can serve as a state-space for inputs confined to \mathfrak{B} . For this purpose let us denote by $S(t)$ the shift semigroup on Σ , mapping Σ into Σ , defined by

$$S(t)u = v, \quad y(s) = \begin{cases} u(s + t) & \text{for } -\infty < s \leq -t, \\ 0 & \text{for } -t < s < 0. \end{cases}$$

We note that Σ can be considered as a subspace of the space of continuous linear functionals on the Banach space of continuous functions $C(-\infty, 0]$.

If we topologize Σ by the corresponding weak-star topology, $S(t)$ is a continuous linear transformation for each t . If we denote by δ the delta function supported at the origin, for any $u(\cdot)$ in Σ the integral

$$\int_{t_0}^t S(t-s)\delta u(s) ds$$

is well defined as a Petti's integral and, in fact, corresponds to the function

$$\begin{aligned} u(t+s) & \text{ for } -t+t_0 < s < 0, \\ 0 & \text{ for } -\infty < s < -t+t_0. \end{aligned}$$

For given $u(\cdot)$ in Σ , and each a , $-\infty < a < +\infty$, the function $u(a+s)$, $-\infty < s < 0$, is an element of Σ . If we denote it by $x(a)$, we have the representation:

$$(14) \quad x(t) = S(t-t_0)x(t_0) + \int_{t_0}^t S(t-s)\delta u(s) ds, \quad t \geq t_0.$$

Next, let

$$v(t) = L(t)x(t).$$

Then it is readily verified that the element $v(t)$ in \mathcal{L} corresponds to the function

$$\int_{-\infty}^t W(t+\Delta, s)u(s) ds \quad \text{a.e.,} \quad \Delta > 0.$$

To relate (14) to the output let us define a linear functional on a nonempty linear subspace of \mathcal{L} by

$$f_0(g) = \lim_{\Delta \rightarrow 0} (1/\Delta) \int_0^\Delta g(s) ds,$$

whenever this limit exists and is finite. Then because of our assumptions on the function $W(t, s)$, we note that for any $x \in \Sigma$, $L(t)x$ belongs to the domain of definition of $f_0(\cdot)$, omitting a set of Lebesgue measure zero (which is independent of x). Omitting this set of measure zero, we can define a family of linear (not necessarily continuous) functionals on Σ by

$$c(t)(x) = f_0(L(t)x), \quad x \in \Sigma.$$

Moreover, we have then that for $x(t)$ defined by (14), the output is given by

$$(15) \quad y(t) = c(t)(x(t)) \quad \text{a.e.}$$

We note that (14) and (15) describe the system in terms of state-input and output-state relations.

We proceed next to enlarge the state-spaces and accommodate all inputs for which the output is defined as measurable functions essentially bounded on finite intervals. We can do this by first reducing the state-spaces and then introducing a new topology—the “output-induced” topology—and completing the space in that topology. For each t , let

$$\Sigma_0(t) = [x \in \Sigma | L(t) x = 0].$$

The “reduced states at time t ” are the elements of the subspace complementary to the subspace $\Sigma_0(t)$. Let us denote the complementary space by

$$\Sigma_R(t).$$

This space is, of course, (algebraically) isomorphic to the factor space modulo $\Sigma_0(t)$. We topologize $\Sigma_R(t)$ by inducing the minimal topology which makes $L(t)$ continuous, considering Σ topologized by the denumerable seminorms

$$P_n(f) = \text{ess sup}_{0 < t < n} |f(t)|$$

for each positive integer n . Since Σ is then a locally convex space, so is $\Sigma_R(t)$ for each t . Let us denote the linear mapping defined on Σ mapping Σ onto $\Sigma_R(t)$. Suppose x is an element of $\Sigma_0(s)$. Then it is readily verified that

$$P(t)S(t-s)x = 0 \quad \text{for } t \geq s.$$

Hence we can define a two-parameter family of linear transformations $T(t; s)$, $-\infty < s \leq t < +\infty$, mapping $\Sigma_R(s)$ into $\Sigma_R(t)$, defined by

$$T(t; s)x = P(t)S(t-s)y,$$

where y is any element in Σ such that

$$P(s)y = x.$$

It will be convenient to write this as

$$T(t; s) = P(t)S(t-s)P(s)^{-1}.$$

It is readily verified that $T(t; s)$ is a linear continuous transformation. Also we have the “transition” property:

$$(16) \quad T(t; s) = T(t; \tau)T(\tau; s), \quad s \leq \tau \leq t.$$

Next let

$$B(t) = P(t)\delta.$$

Then from (14) we can write

$$P(t) x(t) = P(t) S(t - t_0) P(t_0)^{-1} P(t_0) x(t_0) \\ + \int_{t_0}^t P(t) S(t - s) P(s)^{-1} B(s) u(s) ds,$$

so that in terms of reduced states we have the relation

$$(17) \quad \hat{x}(t) = T(t; t_0)x(t_0) + \int_{t_0}^t T(t; s)B(s)u(s) ds,$$

where the $\hat{}$ indicates that the states are reduced at the indicated times. Also, we obviously have in place of (15),

$$(18) \quad \begin{aligned} y(t) &= C(t) [P(t)^{-1} \hat{x}(t)] \quad \text{a.e.} \\ &= C(t) [\hat{x}(t)] \quad \text{a.e.} \end{aligned}$$

It is not difficult to see that relations (17) and (18) can be extended to the completions of the spaces $\Sigma_R(t)$. We shall denote the completed spaces by $\overline{\Sigma_R(t)}$. Thus (17) and (18) are the state-input and output-state relations in terms of reduced states.

Next we shall show that the weak connectedness property holds for the completed reduced states. Thus, let the reduced state $x(t)$ be given, and let us assume first that it is in $\Sigma_R(t)$. Let $u(\cdot)$ be in Σ , such that

$$P(t) u = x(t).$$

Then we know that $u(s)$ vanishes for $s < -t_0$ for some positive t_0 . It is now readily verified that

$$u = \int_{t-t_0}^t S(t-s)\delta u(s-t) ds,$$

and hence we have

$$(19) \quad \hat{x}(t) = \int_{t-t_0}^t T(t; s)B(s)v(s) ds,$$

where

$$v(s) = u(s-t), \quad t-t_0 < s < t,$$

and (19) verifies the required property. If $\hat{x}(t)$ is in the completed space $\overline{\Sigma_R(t)}$, then it is clear from the definition that we can find a sequence $v_n(\cdot)$, such that

$$\hat{x}(t) = \lim \int_{t-t_n}^t T(t; s)B(s)v_n(s) ds.$$

This completes the proof of the theorem.

Let us next consider the spaces $\Sigma_R(t)$. Since these need be only deter-

mined within an isomorphism, we can take

$$(20) \quad \Sigma_R(t) = L(t) \Sigma$$

with the topology as a subspace of \mathfrak{L} . The completion $\Sigma_R(t)$ is simply the closure of the subspace on the right of (20). We now define Σ as the closed linear subspace in \mathfrak{L} generated by the subspaces $\Sigma_R(t)$, $-\infty < t < +\infty$, and, in particular, we can then discuss the continuity of

$$T(t; s)x, \quad x \in \Sigma_R(s),$$

as a function of t . Thus

$$\begin{aligned} T(t + \Delta; s)x - T(t; s)x &= (T(t + \Delta; t) - I) T(t; s)x \\ &= (T(t + \Delta; t) - I)x(t), \quad x(t) \in \Sigma_R(t). \end{aligned}$$

Now for any x in $\Sigma_R(t)$,

$$T(t + \Delta; t)x - x = L(t + \Delta) S(\Delta) u - L(t)u,$$

with $x = L(t)u$ and the element on the right given by

$$\int_{-\infty}^0 (W(t + \Delta, t + s) - W(t, t + s))u(s) ds,$$

$$0 < \sigma < \infty.$$

Since, in particular, we may take $u(\cdot)$ to be a delta function, it follows that $T(t; s)x$ is continuous in t , $t \geq s$, if and only if, for each nonpositive s and each L ,

$$\operatorname{ess\,sup}_{0 < \sigma < L} |W(t + \Delta + \sigma, t + s) - W(t + \sigma, t + s)|$$

goes to zero with Δ . To see that this is a sufficient condition we have also to make use of the fact that $W(t, s)$ is bounded on finite intervals.

Next let us consider the consequences of finite dimensionality of the reduced state-spaces. First of all we shall consider a simple case to show that the situation is not the same as in the time-invariant case. Thus, let $F(\cdot)$ denote a finitely additive set function defined on the field of finite unions of half-open intervals, such that

$$g(t; s) = F([s, t))$$

is measurable in s, t and bounded (from above) on finite intervals. We note that such a function need not be countably additive. Let

$$(21) \quad \begin{aligned} W(t, s) &= \exp g(t, s), & t > s, \\ W(t, t) &= 1. \end{aligned}$$

Then we have

$$(22) \quad W(t, s) = W(t, \tau) W(\tau, s), \quad s < \tau < t.$$

Hence for any u in Σ ,

$$(23) \quad L(t) u = a(t, u) L(t) \delta,$$

where

$$a(t, u) = \int_{-\infty}^0 W(t, s) u(s) ds,$$

and $L(t)\delta$ is the function

$$W(t + \sigma, t), \quad 0 < \sigma < \infty.$$

It follows that $\Sigma_R(t)$ is of dimension at most one for each t . But it can be of dimension zero for some t . For example, we have only to choose $g(\cdot, \cdot)$ so that

$$g(t, s) = -\infty \quad \text{for } s \leq t_0 < t$$

but finite otherwise. Specifically, let

$$(24) \quad W(t, s) = \exp \left(- \int_s^t 1/|x| dx \right).$$

Then $a(t, u)$ is zero for every u for $t > 0$, and hence $\Sigma_R(t)$ is of dimension zero for positive t . One consequence of this is that a system can have finite-dimensional state-spaces and yet need not be dynamic. To see this, let us pursue the same example with $W(t, s)$ given by (24). Suppose the system were dynamic so that the input and output are related by equations of the form:

$$\begin{aligned} \dot{x}(t) &= A(t) x(t) + B(t) u(t), \\ y(t) &= C(t) x(t). \end{aligned}$$

Let $\Phi(t)$ be the fundamental matrix solution of the homogeneous equation. Then it would follow that we must have

$$W(t, s) = C(t) \Phi(t) \Phi(s)^{-1} B(s), \quad s < t.$$

Taking the input $u(t)$ to be zero for $t > 0$, we have

$$\begin{aligned} y(t) &= \int_{-\infty}^0 W(t, s) u(s) ds \\ &= C(t) \Phi(t) \Phi(s)^{-1} x(0) = 0, \quad t > 0. \end{aligned}$$

But $x(0)$ can be chosen arbitrarily here, and for each such choice there

must be WNC, corresponding to an input $u(t)$ in $t \leq 0$. Hence

$$C(t) \Phi(t) \Phi(0)^{-1} x(0) = 0, \quad t > 0.$$

Hence $C(t)$ must be zero for positive t almost everywhere. But this contradicts the fact that

$$0 \neq W(t, s) = C(t) \Phi(t) \Phi(s)^{-1} B(s), \quad 0 < s < t.$$

We can now state some necessary and sufficient conditions for the reduced state-spaces to be finite-dimensional.

THEOREM 3.2. *A necessary and sufficient condition for the system given by the input-output relation (11) to have a reduced state-space of finite dimension at any given time t is that the following relation hold:*

$$(25) \quad W(t + \Delta, t - s) = \sum_{j=1}^n b_j(s; t) W(t + \Delta, t - s_j)$$

for every Δ, s nonnegative and fixed, s_j nonnegative.

Proof. Suppose condition (25) is satisfied; then, for any u in Σ , it follows readily that

$$L(t) u = \sum_{j=1}^n c_j L(t) \delta_j,$$

where δ_j represents the delta function:

$$\delta_j(s) = \delta(s + s_j), \quad s \leq 0,$$

and hence $\Sigma_R(t)$ is of finite dimension. Conversely, suppose $\Sigma_R(t)$ is finite-dimensional. Then let us consider the linear space generated by the elements of the form

$$L(t) \delta_\tau, \quad 0 \leq \tau,$$

where

$$\delta_\tau(s) = \delta(s + \tau), \quad 0 \geq s.$$

Since this space has finite dimension, let $\{L(t) \delta_{\tau_j}\}$ be a basis. Then (25) follows, since $L(t) \delta_\tau$ corresponds to the function

$$W(t + \Delta, t - \tau), \quad 0 < \Delta < \infty.$$

Also it readily follows that $\{L(t) \delta_{\tau_j}\}$ provides a basis for $\Sigma_R(t)$ as well. This completes the proof of the theorem.

Let us note next that in a linear WNC system (whether WNC is defined in terms of states or by (10)), if the reduced state-space $\Sigma_R(t)_0$ is finite-dimensional for some t_0 , then so is $\Sigma_R(t)$ for every $t \geq t_0$. Indeed, the dimension of $\Sigma_R(t)$ is a monotone decreasing function of t . This can be

seen readily as follows. Let $x(t)$ be any element of $\Sigma_R(t)$. Then, by WNC,

$$\begin{aligned} x(t) &= \int_{t-L}^t T(t; s) B(s) u(s) ds \\ &= T(t; t_0) x(t_0) + \int_{t_0}^t T(t; t_0) T(t_0, s) B(s) u(s) ds, \quad t > t_0, \end{aligned}$$

so that $x(t)$ is in the range of $T(t; t_0)$, or the range of $T(t; t_0)$ is precisely $\Sigma_R(t)$, and the range of a linear transformation cannot have a dimension larger than that of its domain space. But, of course, the dimension may be actually less as we have already shown by example. However, we have the following theorem.

THEOREM 3.3. *Suppose a linear system is WNC (whether in terms of states or in terms of input-output as in (10)) and suppose the dimension of the reduced state-space at every instant is exactly the same and finite. Then the system may be represented as a dynamic system.*

Proof. We shall use the notation already developed. Because $\Sigma_R(t)$ is given to be finite-dimensional, and it is a linear topological space in the output induced topology, the completion of $\Sigma_R(t)$ is still $\Sigma_R(t)$ and is isomorphic and homeomorphic to a unitary space of the given dimension, say n . Let $\hat{P}(t)$ denote the 1-1 linear mapping onto the unitary space E_n of dimension n . Let

$$\Phi(t, s) = \hat{P}(t) T(t; s) \hat{P}(s)^{-1}, \quad s \leq t.$$

Then since the range of $T(t; s)$ is precisely $\Sigma_R(t)$, the linear transformation $\Phi(t, s)$ is nonsingular. Hence let us define:

$$(26) \quad \Phi(s, t) = \Phi(t, s)^{-1} \quad \text{for } s \leq t.$$

Then we have

$$(27) \quad \Phi(t, s) = \Phi(t, \tau) \Phi(\tau, s) \quad \text{for any } \tau, \quad -\infty < \tau < +\infty.$$

If we now define:

$$\psi(t) = \Phi(t, a) \quad \text{for some fixed } a,$$

we obtain

$$(28) \quad \Phi(t, s) = \psi(t) \psi(s)^{-1}.$$

Next from (17) we can write the corresponding state-input relation for elements in E_n :

$$(29) \quad x(t) = \Phi(t, t_0) x(t_0) + \int_{t_0}^t \Phi(t, s) \hat{P}(s) B(s) u(s) ds.$$

Again, (18) defines a linear functional on $\Sigma_R(t)$ omitting the exceptional t -set of measure zero. But since $\Sigma_R(t)$ is now finite-dimensional, we must have

$$(30) \quad \dot{c}(t) (\hat{P}(t)^{-1} x) = [C(t), x], \quad x \in E_n,$$

where $[\cdot, \cdot]$ denotes the inner product in E_n , and $C(t)$ is a function defined almost everywhere, with range in E_n , with the property that

$$[C(t), x] = 0 \quad \text{a.e. for } t \geq t_0$$

implies x is zero. Also (30) is essentially bounded on bounded intervals. Thus the output-state relation becomes

$$(31) \quad y(t) = [C(t), x(t)] \quad \text{a.e.,}$$

where $x(t)$ is defined by (29). We claim now that $u(t)$, $y(t)$ may be expressed in terms of the dynamic equations:

$$(32) \quad \dot{z}(t) = \psi(t)^{-1} \hat{P}(t) B(t) u(t),$$

$$(33) \quad y(t) = [C(t), \psi(t) z(t)].$$

For this we have only to use (28) and the fact that the system is WNC in terms of the states. Also, clearly any solution of (32) and (33) can be expressed in the form (29) and (31). The system is thus representable as a dynamic system.

COROLLARY. *Under the conditions of Theorem 3.3, the input-output relation may be expressed as*

$$(34) \quad y(t) = \left[\hat{C}(t), \int_{-\infty}^t D(s) u(s) ds \right],$$

where $\hat{C}(t)$, $D(s)$ are functions (Lesbesgue measurable and essentially bounded on finite intervals) with values in E_n , and the range of the linear transformation

$$(35) \quad \int_{-\infty}^0 D(t+s) u(s) ds,$$

mapping Σ into E_n , is all of E_n for each t . Conversely, any system whose input-output relation may be expressed by (34) is a linear dynamic system, provided (35) holds.

Proof. Under the conditions of the theorem we have, from WNC and (33),

$$W(t, s) = [\hat{C}(t), \psi(t) \psi(s)^{-1} \hat{P}(s) B(s)],$$

from which (34) follows with

$$D(s) = \psi(s)^{-1} \hat{P}(s) B(s), \quad C(t) = \psi(t)^* C(t).$$

Again, for each $u(\cdot)$ in Σ , $L(t)u$ now corresponds to the function

$$\left[C(t + \Delta), \int_{-\infty}^0 D(t + s)u(s) ds \right], \quad \Delta \geq 0,$$

and the range of $L(t)$ is isomorphic to $\Sigma_{\mathcal{R}}(t)$ which has dimension n for each t , by assumption, and hence (35) follows. Conversely, if (34) and (35) hold, we can write

$$(36) \quad \dot{z}(t) = D(t) u(t),$$

$$(37) \quad y(t) = [\hat{C}(t), z(t)],$$

and the system will of course have the WNC property in addition. In fact, (36) has the solution

$$z(t) z(t_0) + \int_{t_0}^t D(s)u(s) ds,$$

and, by (35), we can write

$$z(t_0) = \int_{-\infty}^0 D(t_0 + s)u(s) ds = \int_{-\infty}^{t_0} D(s)u(s - t_0) ds,$$

or (36) and (37) are equivalent to (34).

Remark. The condition that the reduced state-spaces have the same dimension may be replaced by the condition that the reduced state-spaces are finite-dimensional and the transformation $T(t, s)$ is continuous in $t \geq s$ for each s . For, by a result given by Aczel [9], this would imply that $T(t; s)$ is nonsingular for $t \geq s$ and hence (28) may be deduced.

REFERENCES

- [1] L. ZADEH AND C. DESOER, *Linear System Theory*, McGraw-Hill, New York, 1964.
- [2] A. V. BALAKRISHNAN, *Linear systems with infinite-dimensional state-spaces*, Proc. Symposium on System Theory, Brooklyn Polytechnic Institute, New York, 1965, pp. 69-88.
- [3] ———, *On the state-space theory of linear systems*, J. Math. Anal. Appl., 14 (1966), pp. 371-391.
- [4] ———, *On the controllability of a nonlinear system*, Proc. Nat. Acad. Sci. U.S.A., 55 (1966), pp. 465-368.
- [5] R. E. KALMAN, *Mathematical description of linear dynamic systems*, this Journal, 1 (1963), pp. 152-192.
- [6] L. S. PONTYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

- [7] D. C. YOULA, *The synthesis of linear dynamic systems from prescribed weighting functions*, SIAM J. Appl. Math., 14 (1966), pp. 527-549.
- [8] E. G. GILBERT, *Controllability and observability in multivariable control systems*, this Journal, 1 (1963), pp. 128-151.
- [9] J. ACZEL, *Lectures on Functional Equations and their Applications*, Academic Press, New York, 1966.

ON THE PROBLEM OF APPROXIMATE SYNTHESIS OF OPTIMAL CONTROLS*

T. F. BRIDGLAND, JR.†

1. Let E^n denote n -dimensional Euclidean space and let $\{\Omega^n; d\}$ denote the space of nonvoid compact subsets of E^n metrized by the Hausdorff distance, d . Let $f: E^1 \times E^n \times E^m \rightarrow E^n$, $\Phi: E^1 \times E^n \rightarrow \{\Omega^n; d\}$, $G: E^1 \times E^n \rightarrow \{\Omega^n; d\}$ be continuous functions and let $\mathfrak{U}(t_0, x_0)$ denote the set of functions $u: E^1 \rightarrow E^m$ having bounded, Lebesgue measurable components and satisfying

$$(1) \quad u(t) \in \Phi(t, x_u(t; t_0, x_0)),$$

where $x_u(\cdot; t_0, x_0)$ is a solution of

$$(2) \quad \dot{x} = f(t, x, u(t)), \quad x(t_0) = x_0.$$

Under the assumption that there exists a (control) function $u \in \mathfrak{U}(t_0, x_0)$ for which $x_u(t_1; t_0, x_0) \in G(t_1)$, the time-optimal control problem for (2) is that of determining a control function which yields

$$\min \{t_1 \mid x_u(t_1; t_0, x_0) \in G(t_1); u \in \mathfrak{U}(t_0, x_0)\}.$$

As is well known, Filippov [1] has shown under mild conditions, the principal one of which is the convexity of the contingent

$$(3) \quad R(t, x) = \{f(t, x, \phi) \mid \phi \in \Phi(t, x)\},$$

that a solution to the time-optimal control problem exists.

In a recent paper [2], Hermes has examined the problem of approximation of the solution of the time-optimal control problem for (2). Hermes' technique consists of replacing a consideration of this problem by the consideration of a sequence of time-optimal control problems obtained by approximating the contingent (3) by contingents $R^\epsilon(t, x) \supset R(t, x)$ having the properties: (i) the Hausdorff distance between $R^\epsilon(t, x)$ and $R(t, x)$ tends to zero as $\epsilon \rightarrow 0$; (ii) $R^\epsilon(t, x)$ possesses a smooth boundary with positive Gaussian curvature. Hermes shows [2, Theorem 5] that the family of solutions of the approximating problems contains a subsequence converging uniformly to a solution of the time-optimal control problem for (2). Hermes has suggested that the smoothness of the functions involved

* Received by the editors November 17, 1966, and in revised form February 23, 1967.

† University of Alabama Research Institute, Huntsville, Alabama 35807. The research reported in this paper was supported by the National Aeronautics and Space Administration under Grant NsG-381.

in his approximating problems should make these problems accessible to treatment by means of the Carathéodory technique utilized by Kalman [3] to prove existence of optimal feedback controls. However, aside from a brief example, Hermes does not pursue this line of thought.

In the attempt to implement Hermes' suggestion one is confronted with some technical difficulties, inhering to the time-optimal control problem, the resolution of which can be accomplished apparently only by the distasteful expedient of adoption of a plethora of unnatural hypotheses. Rather than follow this course, in this paper we pursue Hermes' suggestion in the context of the minimum miss distance problem [4], a problem which is more fundamental than the time-optimal control problem and the formulation of which is such as to permit a natural fulfillment of the working hypotheses of the Carathéodory method. We shall thus be able to focus our attention on the principal aspects of Hermes' method without being more than minimally distracted by peripheral considerations.

The purpose of this paper is two-fold. First, we wish to give an exposition of Hermes' method which will display at once the essential features of the method and its applicability in contexts broader than that to which Hermes confined his attention. Second, we wish to examine the extent to which Hermes' method is applicable to the problem of establishing existence of optimal feedback controls and to indicate questions relating to this application upon which further research is needed.

To accomplish these purposes, the following format has been adopted for this paper. In §2 we paraphrase Hermes' approximation theorem [2, Theorem 4] in a form which is free of reference to *any* optimal control problem and give a proof, alternative but parallel to Hermes' proof, which displays the features of the approximation which make it attractive for use in connection with the Carathéodory technique. Our proof also permits us to dispense with Hermes' concept of equivalent optimal control problems. In §3 we formulate the minimum miss distance problem precisely and summarize the results of [4] which are pertinent to the present study. The principal result of §3 is Theorem 3, which is the counterpart of [2, Theorem 5]. In §4 we examine the question of existence, and other properties, of solutions of a Hamilton-Jacobi equation involving the type of Hamiltonian function obtained as a natural consequence of our version of Hermes' approximation theorem. Finally, in §5 we discuss the applicability of the results of the preceding sections to the problem of existence of (approximate) syntheses of minimum miss distance controls.

2. Before paraphrasing Hermes' approximation theorem [2, Theorem 4] it will be useful to summarize here some properties of nonvoid, compact, convex subsets of E^n . Let us denote by $\{\Gamma^n; d\}$ the space of all such subsets, metrized by the Hausdorff distance, d . If $A \in \{\Gamma^n; d\}$, the support function,

m , of A is defined on E^n by

$$m(p) = \max \{p \circ \sigma \mid \sigma \in A\};$$

we prefer to work with the dual, h , of m defined on E^n by

$$h(p) = \min \{p \circ \sigma \mid \sigma \in A\}.$$

It is this latter function that we shall refer to as "support function" throughout the remainder of this paper. Of course, a theorem concerning either of these functions has associated with it a dual theorem concerning the other function by virtue of the relation $h(p) = -m(-p)$. Thus, for example, since m is positively homogeneous and subadditive, h is positively homogeneous and superadditive.

LEMMA 1. (i) *If $Q: E^m \rightarrow \{\Gamma^n; d\}$ is continuous, then, with the support function of $Q(y)$ denoted by $h(y, \cdot)$, the function $h(\cdot, \cdot)$ is continuous on $E^m \times E^n$.*

(ii) *If $A \in \{\Gamma^n; d\}$ is strictly convex, then the gradient, h_p , of h exists and satisfies $h(p) = p \circ h_p(p)$ and $h_p(p) \in \partial A$, the boundary of A ; moreover, $\sigma \in A$ and $\sigma \neq h_p(p)$ imply $p \circ \sigma > p \circ h_p(p)$, whence $h_p(\alpha p) = h_p(p)$ for $\alpha > 0$.*

(iii) *If Q satisfies the hypothesis of (i) and, in addition, $Q(y)$ is strictly convex for every y , then the function $h_p(\cdot, \cdot)$ is continuous on $E^m \times E^n$.*

Proof. (i) That $h(\cdot, p)$ is continuous for each fixed $p \in E^n$ is a direct consequence of a result of Bonnesen and Fenchel [5, p. 35]; an obvious modification of a proof of Eggleston [6, Theorem 24] suffices to establish the existence, for each $y_0 \in E^m$, $p_0 \in E^n$, of a neighborhood $N(y_0, p_0)$ of y_0 for which the family $\{h(y, \cdot) \mid y \in N(y_0, p_0)\}$ is equicontinuous at p_0 . With these results, an estimate from the triangle inequality substantiates the assertion.

(ii) This is an easy consequence of the remarks of Eggleston [6, pp. 56-57].

(iii) This assertion follows from (i), (ii) by means of an application of an implicit function theorem to the equation

$$h(y, p) - p \circ \sigma = 0.$$

We paraphrase Hermes' approximation theorem in the following way.

THEOREM 1. *Let $R: E^m \rightarrow \{\Gamma^n; d\}$ be continuous and let \mathcal{D}^* be a compact subset of E^m ; then for each $\epsilon > 0$ there exists a continuous function R^ϵ on \mathcal{D}^* into $\{\Gamma^n; d\}$ satisfying:*

(i) *$R(y) \subset R^\epsilon(y)$ and $d(R(y), R^\epsilon(y)) < \epsilon$ on \mathcal{D}^* , where $d(\cdot, \cdot)$ denotes Hausdorff distance;*

(ii) *the support function, $g^\epsilon(y, \cdot)$, of $R^\epsilon(y)$ has the properties that (a) $g^\epsilon(\cdot, \cdot)$ is of class C^2 on $\mathcal{D}^* \times E^n$ and (b) on $\mathcal{D}^* \times E^n$, $g^\epsilon(y, p) \in \partial R^\epsilon(y)$*

and satisfies both $p \circ g_p^\epsilon(y, p) = g^\epsilon(y, p)$ and, for $\alpha > 0$,

$$g_p^\epsilon(y, \alpha p) = g_p^\epsilon(y, p).$$

Our proof is modeled on that of Hermes, the essential differences arising by virtue of our reliance on the fact [7, p. 62] that sets in $\{\Gamma^n; d\}$ are characterized by their support functions.

Proof. For fixed $\epsilon > 0$ let $\delta = \delta(\epsilon)$ be chosen (and fixed) in such a way that, for $y, y^1 \in E^m$ and $\|y - y^1\| < \delta$,

$$d(R(y), R(y^1)) < \epsilon/8.$$

Letting \mathfrak{N} denote the compact 2δ neighborhood of \mathfrak{D}^* , we apply to \mathfrak{N} the simplicial approximation utilized by Hermes [2, pp. 417–418]. Using Hermes' notation, we denote by \mathfrak{D} the union of members of the family, $\{\bar{\sigma}_g^m\}$, of geometric simplices having all vertices in \mathfrak{N} ; evidently $\mathfrak{D}^* \subset \mathfrak{D}$. By construction, each member of this family has diameter less than δ ; moreover, each point $y \in \mathfrak{D}$ has a unique representation of the form

$$y = \sum_{i=1}^{m+1} \alpha_i(y) \cdot y_i,$$

where each $\alpha_i(\cdot)$ is uniformly Lipschitz continuous on \mathfrak{D} , $0 \leq \alpha_i(y) \leq 1$ and $\sum \alpha_i(y) = 1$, the points y_i being the vertices of the simplex to which y belongs.

Still following Hermes' proof, we associate with each vertex $y_i \in \mathfrak{D}$ a set $Q(y_i, \epsilon)$, which, by virtue of [5, p. 36], may be chosen in such a way that $Q(y_i, \epsilon)$ is compact, strictly convex, possesses an analytic boundary with positive Gaussian curvature, contains an $\epsilon/4$ neighborhood of $R(y_i)$ and whose Hausdorff distance from this neighborhood is less than $\epsilon/4$. Denoting by $h^\epsilon(y_i, \cdot)$ the support function of $Q(y_i, \epsilon)$ (here we digress from Hermes' notation) it then follows from Lemma 1 and [2, Theorem 2] that $h^\epsilon(y_i, \cdot)$ is of class C^2 and that

$$h^\epsilon(y_i, p) = p \circ h_p^\epsilon(y_i, p).$$

At this point we depart from Hermes' line of argument and extend the definition of $Q(\cdot, \epsilon)$ to all of \mathfrak{D} by defining, for each $y \in \mathfrak{D}$,

$$(4) \quad Q(y, \epsilon) = \sum_{i=1}^{m+1} \alpha_i(y) \cdot Q(y_i, \epsilon).$$

Certainly $Q(y, \epsilon)$ is convex for $y \in \mathfrak{D}$. We may now conclude, exactly as does Hermes for his extended function $Q(\cdot, \epsilon)$, that on \mathfrak{D} ,

$$d(R(y), Q(y, \epsilon)) \leq 3\epsilon/4,$$

and $Q(y, \epsilon)$ contains the $\epsilon/8$ neighborhood of $R(y)$. For two points $y^0, y^1 \in \mathfrak{D}$

lying in the same simplex from (4) we obtain

$$Q(y^1, \epsilon) = Q(y^0, \epsilon) + \sum_{i=1}^{m+1} [\alpha_i(y^1) - \alpha_i(y^0)]Q(y_i, \epsilon),$$

and from this equation the continuity of $Q(\cdot, \epsilon)$ on \mathfrak{D} follows readily from the continuity of the functions $\alpha_i(\cdot)$.

If the definition of $h^\epsilon(\cdot, p)$ be extended to \mathfrak{D} by

$$(5) \quad h^\epsilon(y, p) = \sum_{i=1}^{m+1} \alpha_i(y) \cdot h^\epsilon(y_i, p),$$

then it is an immediate consequence of the definitions involved that $h^\epsilon(y, \cdot)$ is the support function of $Q(y, \epsilon)$ for each $y \in \mathfrak{D}$. A point $\sigma \in Q(y, \epsilon)$ has the representation

$$\sigma = \sum_{i=1}^{m+1} \alpha_i(y) \sigma_i,$$

and with this it follows from (5) that $p \circ \sigma = h^\epsilon(y, p)$ if and only if $p \circ \sigma_i = h^\epsilon(y_i, p)$, $i = 1, \dots, m+1$. Hence $Q(y, \epsilon)$ is strictly convex for each $y \in \mathfrak{D}$ and now Lemma 1(ii), 1(iii) permit the conclusion that $h_p^\epsilon(y, \cdot)$ exists and satisfies

$$h^\epsilon(y, p) = p \circ h_p^\epsilon(y, p),$$

and that $h^\epsilon(\cdot, \cdot)$ is continuous. From (5) we obtain further

$$(6) \quad h_p^\epsilon(y, p) = \sum_{i=1}^{m+1} \alpha_i(y) \cdot h_p^\epsilon(y_i, p),$$

which establishes the continuity of $h_p^\epsilon(\cdot, \cdot)$ as well as the fact that $h^\epsilon(y, \cdot)$ is of class C^2 . Similarly, from (6) there may be obtained by differentiation a result from which one infers without difficulty the continuity of $h_{pp}^\epsilon(\cdot, \cdot)$.

Applying to $h^\epsilon(\cdot, \cdot)$ the smoothing technique adopted by Hermes [2, p. 420] we define

$$(7) \quad h^\epsilon(y, p; k) = \int_{E^m} S^k(y - \eta) h^\epsilon(\eta, p) d\eta,$$

where $S^k(\cdot)$ is a mollifier [2, loc. cit.] and $h^\epsilon(\cdot, \cdot)$ is extended to $E^m \times E^n$ by defining $h^\epsilon(\cdot, p)$ to be the zero function on the complement of \mathfrak{D} . Then for every positive integer k , $h^\epsilon(\cdot, \cdot; k)$ is of class C^2 and satisfies

$$h^\epsilon(y, p; k) = p \circ h_p^\epsilon(y, p; k)$$

and

$$h_p^\epsilon(y, \alpha p; k) = h_p^\epsilon(y, p; k), \quad \alpha > 0.$$

Moreover, $h^\epsilon(\cdot, \cdot; k)$ and its derivatives with respect to p tend uniformly on $\mathfrak{D} \times \mathfrak{E}$ with increasing k to $h^\epsilon(\cdot, \cdot)$ and its derivatives with respect to p , where \mathfrak{E} is an arbitrary compact subset of E^n .

It is easy to see that $h^\epsilon(y, \cdot; k)$ is positively homogeneous and super-additive; hence (see [7, Theorem 5.6]) the set $R^k(y; \epsilon)$, defined by

$$(8) \quad R^k(y; \epsilon) = \{z \in E^m \mid z \circ p \geq h^\epsilon(y, p; k) \text{ for all } p \in E^n - \{0\}\},$$

is compact, convex and has $h^\epsilon(y, \cdot; k)$ as its support function so that $R^k(\cdot; \epsilon)$ is continuous. Now $h_p^\epsilon(y, p_0; k) \in \partial R^k(y; \epsilon)$, $p_0 \in E^n$, for by differentiating (7) one may then obtain the result

$$h_p^\epsilon(y, p_0; k) \circ p - h^\epsilon(y, p; k) \geq 0, \quad p \in E^n - \{0\},$$

from the corresponding result for the sets $Q(y, \epsilon)$. There is no difficulty in showing that one can choose k sufficiently large that $R(y) \subset R^k(y; \epsilon)$ and

$$d(R(y), R^k(y; \epsilon)) < \epsilon;$$

for such a choice of k we define

$$R^\epsilon(y) = R^k(y; \epsilon), \quad g^\epsilon(y, p) = h^\epsilon(y, p; k),$$

and the proof is complete.

Remark 1. Note that in the proof of this theorem, once \mathfrak{N} has been defined for a given ϵ , say $\bar{\epsilon}$, the same \mathfrak{N} may be used for all positive $\epsilon < \bar{\epsilon}$. Also, as a consequence of part (i) of the theorem and the result of Bonnesen and Fenchel cited in the proof of Lemma 1, $g^\epsilon(\cdot, \cdot)$ tends uniformly on $\mathfrak{D}^* \times \mathfrak{E}$ to the support function of $R(y)$ as $\epsilon \rightarrow 0$.

Remark 2. There follows from Theorem 1 the fact that $R(y)$ can always be approximated by a decreasing sequence

$$R^{\epsilon_1}(y) \supset R^{\epsilon_2}(y) \supset \cdots \supset R(y)$$

on \mathfrak{D}^* . This is accomplished, for example, by taking $\epsilon_n = 2^{-n}$, $n = 0, 1, 2, \dots$, and choosing R^{ϵ_n} so that $R^{\epsilon_n}(y)$ contains the closed $2^{-(n+1)}$ neighborhood of $R(y)$ and is contained in the 2^{-n} neighborhood of $R(y)$.

3. Now let us formulate the minimum miss distance problem and summarize the results of [4] concerning this problem. Denoting by I_T the compact interval $\{t \mid 0 \leq t \leq T\}$, for the functions f, Φ, G introduced in §1 we shall consider their restrictions to the respective domains $I_T \times E^n \times E^m$, $I_T \times E^n$ and $I_T \times E^n$. We continue to assume the continuity of the functions so restricted and we require further:

- (i) the value of Φ is constant on $I_T \times E^n$ and for each $u \in \mathfrak{U}(t_0, x_0)$, (2) has a unique solution;
- (ii) the set $R(t, x)$ defined in (3) is convex for each $(t, x) \in I_T \times E^n$;
- (iii) there exists $C > 0$ such that for all $(t, x) \in I_T \times E^n$ and each

$$\sigma \in R(t, x),$$

$$|x \circ \sigma| \leq C(\|x\|^2 + 1).$$

For a point $x \in E^n$ and a nonvoid compact set $A \subset E^n$, the distance $\alpha(x, A)$ between x and A is defined by

$$(9) \quad \alpha(x, A) = \min \{ \|x - a\| \mid a \in A \}.$$

Condition (iii) ensures the continuability to I_T of the solutions of (2) and so a *miss distance*, $\delta(t, x, u)$, may be defined for (2) by

$$(10) \quad \delta(t, x, u) = \min \{ \alpha(x_u(\tau; t, x), G(\tau)) \mid t \leq \tau \leq T \};$$

the *first time of closest approach*, $t^*(t, x, u)$, is defined by

$$(11) \quad t^*(t, x, u) = \min \{ \tau \in [t, T] \mid \alpha(x_u(\tau; t, x), G(\tau)) = \delta(t, x, u) \}.$$

As pointed out in [4], under conditions (i), (ii), δ and t^* are defined respectively by (10), (11) for all $(t, x) \in I_T \times E^n$ and all $u \in \mathfrak{U}(t, x)$. For a point $(t, x) \in I_T \times E^n$, the *set of admissible controls*, $U(t, x)$, is defined by

$$U(t, x) = \{ u \in \mathfrak{U}(t, x) \mid t < t^*(t, x, u) \leq T \};$$

then the *set of admissible initial data*, B_∞ , is defined as

$$B_\infty = \{ (t, x) \in I_T \times E^n \mid U(t, x) \neq \emptyset \},$$

where \emptyset denotes the null set. We may now state our final working hypothesis, which is simply:

(iv) $B_\infty \neq \emptyset$.

The intuitive meaning of (iv) is that there exists at least one point (t_0, x_0) from which a "closer approach" to the target is possible along some trajectory of (2).

As defined in [4], the minimum miss distance problem is that of determining for each $(t, x) \in B_\infty$ a control $u^0 \in U(t, x)$ which yields

$$(12) \quad \min \{ \delta(t, x, u) \mid u \in U(t, x) \}.$$

Part of the content of [4, Theorem 4] is the assertion that conditions (i)-(iv) imply the existence of such a control. A formulation, implicit in [4], of the minimum miss distance problem which is equivalent to (12) will better suit our present purpose. In order to give this formulation, we must first define the *kernel*, $K(t, x)$, of the set of admissible controls:

$$K(t, x) = \{ u \in U(t, x) \mid (t^*(t, x, u), x_u(t^*(t, x, u); t, x)) \notin B_\infty \}.$$

The assertion of [4, Theorem 3] is that condition (iv) implies that $K(t, x) \neq \emptyset$ for every $(t, x) \in B_\infty$. Then, as indicated in the proof of [4, Theorem 4], (12) is equivalent to

$$(13) \quad \min \{ \delta(t, x, u) \mid u \in K(t, x) \};$$

this is the formulation of the minimum miss distance problem with which we shall work in the sequel.

With B_∞ are associated sets B_γ defined by

$$B_\gamma = \{ (t, x) \in B_\infty \mid \delta(t, x, u) < \gamma \text{ for some } u \in U(t, x) \}, \quad 0 < \gamma, \\ B_0 = \bigcap_{0 < \gamma} B_\gamma.$$

In [4], the following properties of B_γ , $0 < \gamma \leq \infty$, were shown to obtain under conditions (i)-(iv):

(a) B_0 is the (possibly empty) set of points (t_0, x_0) from which some trajectory of (2) "hits" the target;

(b) $B_\infty = \bigcup_{0 < \gamma} B_\gamma$ and, for $0 < \gamma \leq \infty$, B_γ is relatively open in $I_T \times E^n$;

(c) for $\gamma > 0$, B_γ is bounded.

In fact, for $\gamma > 0$ and $B_\gamma \neq \emptyset$, we have

$$(14) \quad B_\gamma \subset \{ (t, x) \in I_T \times E^n \mid \|x\|^2 < [(\gamma + \beta)^2 + 1] \exp 2CT - 1 \},$$

where

$$\beta = \max_{t \in I_T} \{ \max_{\zeta \in G(t)} \|\zeta\| \}.$$

By means of an estimate from condition (iii) there is thus obtained the following lemma.

LEMMA 2. If $B_\gamma \neq \emptyset$ then for each $(t_0, x_0) \in B_\gamma$ and all $u \in \mathfrak{U}(t_0, x_0)$,

$$1 + \|x_u(t; t_0, x_0)\|^2 \leq [(\gamma + \beta)^2 + 1] \exp 4CT, \quad t_0 \leq t \leq T.$$

Consequently, if $\mathfrak{D}^* \subset I_T \times E^n$ be defined by

$$(15) \quad \mathfrak{D}^* = \{ (t, x) \in I_T \times E^n \mid \\ \|x\| \leq [((\gamma + \beta)^2 + 1) \exp (4C + 1)T - 1]^{1/2} \},$$

we see that every integral curve of (2) passing through a point $(t_0, x_0) \in B_\gamma$ remains in \mathfrak{D}^* to the right of t_0 . We suppose henceforth that a $\gamma > 0$ has been chosen, and fixed, for which $B_\gamma \neq \emptyset$; the existence of such a γ is not in doubt in view of property (b) above.

If, as is readily justified, E^m be replaced in Theorem 1 by $E^1 \times E^n$, if the \mathfrak{D}^* of that theorem be that defined in (15) and if the function R be that defined in (3), then in view of condition (ii) of this section the conclusion of Theorem 1 obtains. Thus for each $\epsilon > 0$ we have an approximant $R^\epsilon(t, x)$ to $R(t, x)$ on \mathfrak{D}^* with support function $g^\epsilon(t, x, \cdot)$.

Let us define a set \mathfrak{B} by

$$(16) \quad \mathfrak{B} = \overline{\mathfrak{D}^* - (\mathfrak{D}^* \cap B_\infty')},$$

where B_∞' is the complement of B_∞ in $I_T \times E^n$ and the superior bar denotes closure; evidently $B_\gamma \subset \mathfrak{B}$ so that \mathfrak{B} is bounded and has a nonvoid interior. If $z: I_T \rightarrow E^n$ is continuous and $(t_0, z(t_0)) \in \mathfrak{B}$ for some t_0 , then since B_∞' is relatively closed in $I_T \times E^n$ it is easy to see that there is a smallest $\lambda > t_0$ such that $z(\lambda) \in \mathfrak{D}^* \cap B_\infty'$; we denote this λ by $\lambda(z)$. For a point $(t_0, x_0) \in \mathfrak{B}$, the set $P_\epsilon(t_0, x_0)$ of ϵ -admissible functions is defined to be the collection of all continuous functions $z: I_T \rightarrow E^n$ satisfying:

- (F₁) z is absolutely continuous and $z(t_0) = x_0$;
- (F₂) $(t, z(t)) \in \mathfrak{B}$ for all $t \in [t_0, \lambda(z)]$;
- (F₃) $\dot{z}(t) \in R^\epsilon(t, z(t))$ for almost all $t \in [t_0, T]$.

Since $R(t, x) \subset R^\epsilon(t, x)$ on \mathfrak{B} , it follows that $P_\epsilon(t_0, x_0) \neq \emptyset$ for $(t_0, x_0) \in B_\gamma$; the truth of this assertion is immediately clear upon choosing $u \in K(t_0, x_0)$ and invoking Lemma 2. Moreover, Hermes' argument [2, p. 422]—in which there are some misprints—shows that if $(t_0, x_0) \in B_\gamma$ and $\epsilon \leq 1$ then (F₃) implies (F₂). *Throughout the remainder of this paper we shall assume $\epsilon \leq 1$.*

The proof of the next theorem is essentially that sketched by Filippov for his generalization [1, pp. 81–82] of [1, Theorem 1] and will therefore be omitted.

THEOREM 2. *If conditions (i)–(iv) are satisfied, then for each $(t_0, x_0) \in B_\gamma$ there exists $z^0 \in P_\epsilon(t_0, x_0)$ which yields*

$$(18) \quad \min \{ \alpha(z(\lambda(z))), G(\lambda(z)) \} \mid z \in P_\epsilon(t_0, x_0) \}.$$

If we denote by $V(t_0, x_0)$, $V^\epsilon(t_0, x_0)$ the values of the expressions in (13) and (18) respectively, we may state the following counterpart of [2, Theorem 5]. Taking into account Remark 2, the proof of the latter theorem is applicable with only minor modifications and will be omitted.

THEOREM 3. *Let conditions (i)–(iv) be satisfied and for $(t_0, x_0) \in B_\gamma$ and $\epsilon_n = 2^{-n}$, $n = 0, 1, 2, \dots$, let $\{z^n(\cdot; t_0, x_0)\}$ denote a sequence of those functions $z \in P_{\epsilon_n}(t_0, x_0)$ whose existence is guaranteed by Theorem 2; then this sequence contains a subsequence which converges uniformly on $[t_0, T]$ to a limit function $z^*(\cdot; t_0, x_0)$ having the following properties:*

(i) $z^*(\cdot; t_0, x_0)$ is absolutely continuous and there exists $u \in K(t_0, x_0)$ for which

$$\dot{z}^*(t; t_0, x_0) = f(t, z^*(t; t_0, x_0), u(t))$$

almost everywhere on $[t_0, T]$;

(ii) $V(t_0, x_0) = \lim_{n \rightarrow \infty} V^{\epsilon_n}(t_0, x_0) = \alpha(z^*(t^*(t_0, x_0, u); t_0, x_0), G(t^*(t_0, x_0, u)))$.

Remark 3. As is the case with [2, Theorem 5], the only properties of $R^\epsilon(t, x)$ used in establishing Theorems 2 and 3 are, in addition to convexity, those listed in Theorem 1(i) and in Remark 2. Hence both theorems re-

main valid for much more general approximants; for example, if one sets $r_\kappa = 2^n$, then the r_n -convex hull $S_{r_n}(t, x)$ of $R(t, x)$ (see [8]) is a suitable approximant for use in connection with Theorems 2 and 3. The value of $S_{r_n}(t, x)$ as approximant arises from the simplicity of its structure—an intersection of r_n -balls—and the fact that the continuity of $S_{r_n}(\cdot, \cdot)$, as well as that of its support function, can be demonstrated directly without recourse to the simplicial approximation of the domain space used in Theorem 1.

Remark 4. Theorem 3 and its counterpart [2, Theorem 5] are almost entirely of theoretical interest since the approximations they yield satisfy only

$$\dot{z}^n(t; t_0, x_0) \in R^{\epsilon_n}(t, z^n(t; t_0, x_0)),$$

whereas, from a practical viewpoint, what is needed is an approximation z^n which satisfies

$$\dot{z}^n(t; t_0, x_0) \in R(t, z^n(t; t_0, x_0)).$$

The attainment of an approximation satisfying the latter condition is dependent upon the existence of a satisfactory solution to the problem of approximate synthesis as will be shown in §5.

4. In order to examine the question of approximate synthesis we must restrict our problem still further by requiring a certain amount of smoothness for the set $\partial B_\infty \cap B_\infty'$ and the function $\alpha(\cdot, G(\cdot))$. We assume:

(v) there exists $\psi: E^n \rightarrow I_T$ which is positive valued and of class C^2 and which satisfies

$$\partial B_\infty \cap B_\infty' \equiv \mathfrak{M} = \{(t, x) \in I_T \times E^n \mid t = \psi(x)\}.$$

With this assumption it follows readily that the set \mathfrak{B} defined in (16) may be represented by

$$\mathfrak{B} = \{(t, x) \in \mathfrak{D}^* \mid t \leq \psi(x)\}.$$

\mathfrak{M} is an n -manifold of class C^2 and the vector $(1, -\psi_x(x))^T$ is an exterior normal to \mathfrak{B} at the point $(\psi(x), x) \in \mathfrak{B} \cap \mathfrak{M}$. For $\alpha(\cdot, G(\cdot))$ we assume:

(vi) the function $\omega: I_T \times E^n \rightarrow E^1$ defined by

$$\omega(t, x) = \alpha(x, G(t))$$

is of class C^2 on a neighborhood of \mathfrak{M} .

The first question we shall investigate in this section is that of the existence and uniqueness of solutions of the Hamilton-Jacobi equation

$$(20a) \quad V_t + g^e(t, x, V_x) = 0$$

under the boundary condition

$$(20b) \quad V(\sigma, \xi) = \omega(\sigma, \xi), \quad (\sigma, \xi) \in \mathfrak{B} \cap \mathfrak{M},$$

where $g^\epsilon(t, x, \cdot)$ is the support function of the approximant $R^\epsilon(t, x)$ of Theorem 1. Let us denote by X the projection of \mathfrak{D}^* on E^n and by ϑ , the maximum of $\|\psi_x(x)\|$ on the compact set X . We may state the following theorem.

THEOREM 4. *If, in addition to conditions (i)-(vi), it be assumed that (vii) there exists ν , $0 < \nu < 1$, such that*

$$\vartheta \|\sigma\| \leq 1 - \nu$$

for all $\sigma \in \partial R(\psi(x), x)$ and all $x \in X$, then for each positive¹ $\epsilon \leq \nu\vartheta^{-1}$ there exists, on a neighborhood \mathfrak{F}^ϵ of $\mathfrak{B} \cap \mathfrak{M}$, a unique solution, W^ϵ , for (20) of class C^2 .

Proof. We need show only that (vii) implies that the initial data $\omega(\psi(\xi), \xi)$ are noncharacteristic, for if this is the case the theorem is a corollary of [9, Theorem 9.1, p. 137]. To this end, let us define

$$\gamma^\epsilon(x) = \sup \{ \|g_p^\epsilon(\psi(x), x, p)\| \mid p \in E^n - \{0\} \};$$

since $\gamma^\epsilon(x)$ also may be expressed as

$$\gamma^\epsilon(x) = \sup \{ \|g_p^\epsilon(\psi(x), x, p/\|p\|)\| \mid p \in E^n - \{0\} \},$$

it is clear that the supremum is a maximum, and then easy estimates show that $\gamma^\epsilon(\cdot)$ is continuous on X . Moreover, if $\zeta(x) \in \partial R(\psi(x), x)$ be a point nearest $\gamma^\epsilon(x)$, then $\|\gamma^\epsilon(x) - \zeta(x)\| < \epsilon$ on X and from this inequality and (vii) one obtains easily

$$(21) \quad \|\psi_x(x)\| \gamma^\epsilon(x) < 1, \quad x \in X,$$

provided $\epsilon \leq \nu\vartheta^{-1}$. By virtue of (21) and the continuity of the functions involved, we may define a positive number, μ^ϵ , by

$$(22) \quad \mu^\epsilon = \max_{\xi \in X} |\omega_t(\psi(\xi), \xi) + g^\epsilon(\psi(\xi), \xi, \omega_x(\psi(\xi), \xi)) \mid (1 - \gamma^\epsilon(\xi) \|\psi_x(\xi)\|)^{-1}.$$

It is a consequence of (21), (22) and Banach's fixed-point theorem that the equation

$$(23) \quad \lambda - g^\epsilon(\psi(\xi), \xi, \lambda\psi_x(\xi) + \omega_x(\psi(\xi), \xi)) = \omega_t(\psi(\xi), \xi)$$

has a unique solution $\lambda = \lambda^\epsilon(\xi)$ in the interval $[-\mu^\epsilon, \mu^\epsilon]$ for each $\xi \in X$. We may now define a function ρ^ϵ on X by

$$(24) \quad \rho^\epsilon(\xi) = \lambda^\epsilon(\xi)\psi_x(\xi) + \omega_x(\psi(\xi), \xi).$$

For the system of equations

$$(25) \quad \begin{aligned} \rho - g^\epsilon(\psi(\xi), \xi, \rho)\psi_x(\xi) &= \omega_t(\psi(\xi), \xi)\psi_x(\xi) + \omega_x(\psi(\xi), \xi), \\ \nu + g^\epsilon(\psi(\xi), \xi, \rho) &= 0 \end{aligned}$$

¹ If $\vartheta = 0$, no upper bound on ϵ is required other than that assumed previously.

with unknown functions ρ , ν , the Jacobian determinant has the value

$$\begin{aligned}
 (26) \quad \det \left(\begin{array}{c|c} I_n - g_p^\epsilon(\psi(\xi), \xi, \rho) \psi_x^T(\xi) & g_p^\epsilon(\psi(\xi), \xi, \rho) \\ \hline 0 & 1 \end{array} \right) \\
 = \det \left(\begin{array}{c|c} I_n & g_p^\epsilon(\psi(\xi), \xi, \rho) \\ \hline \psi_x^T(\xi) & 1 \end{array} \right) \\
 = 1 - g_p^\epsilon(\psi(\xi), \xi, \rho) \psi_x(\xi),
 \end{aligned}$$

where I_n is the $n \times n$ unit matrix. If the value of $\rho^\epsilon(\xi)$, as determined by (23), (24), be substituted for ρ in (26), it then follows from (21) and the Cauchy-Schwarz inequality that the Jacobian determinant of (25) does not vanish for $\xi \in X$. Hence, the solution of (25) given by (24) and

$$\nu^\epsilon(\xi) = \omega_\epsilon(\psi(\xi), \xi) - \lambda^\epsilon(\xi)$$

is unique and of class C^1 on X . Thus the initial data $\omega(\psi(\xi), \xi)$ are noncharacteristic and the proof is complete.

For purposes of completeness, we outline the essentials of the proof of [9, Theorem 9.1, p. 137] for the special case treated by Theorem 4. The nontrivial parts of the characteristic differential equations for (20a) are

$$\begin{aligned}
 (27a) \quad \dot{x} &= g_p^\epsilon(t, x, p), \\
 \dot{p} &= -g_x^\epsilon(t, x, p);
 \end{aligned}$$

we denote by $(x^\epsilon(\cdot, \xi), p^\epsilon(\cdot, \xi))^T$ the solution of (27a) satisfying

$$\begin{aligned}
 (27b) \quad x(\psi(\xi), \xi) &= \xi, \\
 p(\psi(\xi), \xi) &= \rho^\epsilon(\xi).
 \end{aligned}$$

The functions $x^\epsilon(\cdot, \cdot)$, $p^\epsilon(\cdot, \cdot)$ are of class C^2 on a neighborhood of $\mathfrak{B} \cap \mathfrak{M}$, so we may consider a transformation

$$(28) \quad \Gamma_\epsilon : \begin{pmatrix} \xi \\ t \end{pmatrix} \rightarrow \begin{pmatrix} x^\epsilon(t, \xi) \\ t \end{pmatrix}$$

on this neighborhood. The Jacobian determinant at $(\psi(\xi), \xi)$ of Γ_ϵ is given by the first expression in (26) with ρ replaced by $\rho^\epsilon(\xi)$, so we may assert the existence of a transformation

$$(29) \quad \Gamma_\epsilon^{-1} : \begin{pmatrix} x \\ t \end{pmatrix} \rightarrow \begin{pmatrix} \xi^\epsilon(t, x) \\ t \end{pmatrix}$$

inverse to Γ_ϵ on a neighborhood \mathfrak{F}_ϵ of $\mathfrak{B} \cap \mathfrak{M}$. On this neighborhood the solution W^ϵ of (20) is given by

$$(30) \quad W^\epsilon(t, x) = \omega(\psi(\xi^\epsilon(t, x)), \xi^\epsilon(t, x)),$$

as may be verified.

COROLLARY (Carathéodory). *If conditions (i)–(vii) are satisfied, then for each positive $\epsilon \leq \nu\vartheta^{-1}$,*

$$W^\epsilon(t, x) = V^\epsilon(t, x), \quad (t, x) \in B_\gamma \cap \mathfrak{F}_\epsilon.$$

Proof. On \mathfrak{F}_ϵ denote by $\bar{x}^\epsilon(\cdot; t_0, x_0)$ the solution of

$$(31) \quad \dot{x} = g_p^\epsilon(t, x, W_x^\epsilon(t, x)), \quad x(t_0) = x_0.$$

It is easy to see both that $B_\gamma \cap \mathfrak{F}_\epsilon \neq \emptyset$ and that $\bar{x}^\epsilon(\cdot; t_0, x_0) \in P_\epsilon(t_0, x_0)$, for in the latter the comments preceding Theorem 2 apply. In the former, if $(t_0, x_0) \in B_\gamma$ and we choose u to be a minimum miss control, then there exists $t \in (t_0, t^*(t_0, x_0, u))$ such that $(t, x_u(t; t_0, x_0)) \in \mathfrak{B} \cap \mathfrak{F}_\epsilon$; but clearly this point is also in B_γ . Defining \bar{t}^ϵ by

$$(32) \quad \bar{t}^\epsilon(t_0, x_0) = \lambda(\bar{x}^\epsilon(\cdot; t_0, x_0)),$$

one deduces from (20) the relation

$$(33) \quad W^\epsilon(t_0, x_0) = \omega(\bar{t}^\epsilon(t_0, x_0), \bar{x}^\epsilon(\bar{t}^\epsilon(t_0, x_0); t_0, x_0)).$$

If z be any other function in $P_\epsilon(t_0, x_0)$, then $W^\epsilon(\cdot, z(\cdot))$ is absolutely continuous and, by virtue of (20a),

$$\frac{d}{dt} W^\epsilon(t, z(t)) = W_x^\epsilon(t, z(t)) \circ \dot{z}(t) - g^\epsilon(t, z(t), W_x^\epsilon(t, z(t)))$$

almost everywhere on $[t_0, \lambda(z))$, since $(t, z(t)) \in \mathfrak{B} \cap \mathfrak{F}_\epsilon$ on this interval. Hence from this result and (F₃), there follows

$$\frac{d}{dt} W^\epsilon(t, z(t)) \geq 0$$

almost everywhere on $[t_0, \lambda(z))$. Integrating this inequality and applying (20b) we obtain finally

$$\omega(\lambda(z), z(\lambda(z))) \geq W^\epsilon(t_0, x_0),$$

which establishes the corollary.

5. From the corollary to Theorem 4 one would like to conclude that (31) furnishes a family of satisfactory approximations to an optimal trajectory of problem (13). That such a conclusion is not necessarily valid is clear upon noting that none of our conditions appears to prevent the diameter of $\mathfrak{B} \cap \mathfrak{F}_\epsilon$ from tending to zero with ϵ . Thus a point (t_0, x_0) may lie in $B_\gamma \cap \mathfrak{F}_{\epsilon_1}$ but not lie in $B_\gamma \cap \mathfrak{F}_{\epsilon_2}$ for some $\epsilon_2 < \epsilon_1$, in which case one must be content with Theorem 3. We are thus presented with one of the major impediments to the application of Hermes' method to the synthesis problem; i.e., lack of sufficiently general conditions ensuring a common domain of existence, for all $\epsilon \leq \nu\vartheta^{-1}$, for the solutions of (20). Conditions ensuring a common

domain of existence are known (see [10]–[13]) for the case in which the manifold \mathfrak{M} of condition (v) is a hyperplane $\{(t, x) \in I_T \times E^n \mid t = \rho\}$, but these conditions involve a knowledge of bounds on the partial derivatives of g^ϵ which are not readily determined for the approximants constructed in the proof of Theorem 1.

Under the assumption that there is a common domain of existence for the functions W^ϵ , the existence of approximations of the type discussed in Remark 4 may be demonstrated. The precise assertion in this connection is contained in Proposition 1 below; in order to prove this proposition we require the following lemma.

LEMMA 3. *Let $Q: I_T \times E^n \rightarrow \{\Gamma^n; d\}$ be continuous and for $q \in E^n$ let $y(t, x, q)$ denote the unique point² in $Q(t, x)$ nearest q ; then the function $y(\cdot, \cdot, \cdot)$ is continuous on $I_T \times E^n \times E^n$.*

Proof. Let (t, x, q) be the limit, as $i \rightarrow \infty$, of a sequence $\{(t^i, x^i, q^i)\}$. Since the function $\alpha(\cdot, \cdot)$ of (9) is continuous on $E^n \times \{\Gamma^n; d\}$ (cf. [14, Lemma 1]) and

$$\alpha(q^i, Q(t^i, x^i)) = \|q^i - y(t^i, x^i, q^i)\|,$$

it follows that if \bar{y} is a limit point, necessarily in $Q(t, x)$, of $\{y(t^i, x^i, q^i)\}$, then

$$\|q - \bar{y}\| = \alpha(q, Q(t, x)).$$

But then \bar{y} is a point nearest q , and since this is unique, continuity of $y(\cdot, \cdot, \cdot)$ follows.

Before stating Proposition 1 it will be convenient to condense some of our previous notation. On the set \mathcal{F}_ϵ of Theorem 4 we define a function k^ϵ by

$$(34) \quad k^\epsilon(t, x) = g_x^\epsilon(t, x, W_x^\epsilon(t, x));$$

on this same set the function \mathcal{K}^ϵ is defined to be the function whose value at (t, x) is the unique point in (the boundary of) $R(t, x)$ nearest $k^\epsilon(t, x)$. For the sequence $\{\epsilon_n\}$ with $\epsilon_n = 2^{-n}$, $n = 0, 1, 2, \dots$, we denote by $\bar{x}^n(\cdot; t_0, x_0)$ the solution of

$$(35) \quad \dot{x} = k^{\epsilon_n}(t, x), \quad x(t_0) = x_0.$$

We note the fact that local existence of solutions of the differential equation

$$(36) \quad \dot{y} = \mathcal{K}^{\epsilon_n}(t, y), \quad y(t_0) = x_0,$$

is assured by the continuity of \mathcal{K}^ϵ which in turn follows from Lemma 3 and the continuity of R and k^ϵ .

² The uniqueness of $y(t, x, q)$ derives from Motzkin's theorem (see [7, Theorem 7.8]).

PROPOSITION 1. *In addition to conditions (i)–(vii) assume that there exists $\bar{\epsilon}$, $0 < \bar{\epsilon} \leq \nu\vartheta^{-1}$, for which $\mathfrak{F}_{\bar{\epsilon}} \subset \mathfrak{F}_{\epsilon}$ for all $\epsilon \leq \bar{\epsilon}$, and let $\bar{t}(t_0, x_0)$ denote*

$$\liminf_{n \rightarrow \infty} \bar{t}^n(t_0, x_0)$$

for $(t_0, x_0) \in B_\gamma \cap \mathfrak{F}_{\bar{\epsilon}}$; then there exists a sequence $\{\bar{x}^n(\cdot; t_0, x_0)\}$ of functions having the following properties:

(i) $\{\bar{x}^n(\cdot; t_0, x_0)\}$ converges uniformly on $[t_0, \bar{t}(t_0, x_0)]$ to a function $z^*(\cdot; t_0, x_0)$ possessing properties (i), (ii) of Theorem 3, where $t^*(t_0, x_0, u) \leq \bar{t}(t_0, x_0)$;

(ii) for each value of n ,

$$\xi^n(t; t_0, x_0) \in \partial R(t, \xi^n(t; t_0, x_0))$$

for $t \in [t_0, \bar{t}(t_0, x_0)]$.

Proof. From the sequence $\{\bar{x}^n(\cdot; t_0, x_0)\}$ defined above we eliminate those initial terms for which $2^{-n} > \bar{\epsilon}$; from what remains we retain only those terms for which n is sufficiently large that $\bar{x}^n(\cdot; t_0, x_0)$ is continuable, as a solution of (35), to the interval $[t_0, \bar{t}(t_0, x_0)]$. That this continuation is possible for large n is a consequence of Theorem 4. The uniform boundedness and equicontinuity on $[t_0, \bar{t}(t_0, x_0)]$ of the resulting sequence may be established in exactly the same way as in the proof of [2, Theorem 5]. Hence there is a subsequence, which we again denote by $\{\bar{x}^n(\cdot; t_0, x_0)\}$, which converges uniformly on $[t_0, \bar{t}(t_0, x_0)]$ to a function $z^*(\cdot; t_0, x_0)$ having properties (i), (ii) of Theorem 3. For the right-hand members of the differential equations (35), of which the $\bar{x}^n(\cdot; t_0, x_0)$ are solutions, we have as a consequence of Theorem 1,

$$(37) \quad \|k^{\epsilon_n}(t, x) - \mathfrak{K}^{\epsilon_n}(t, x)\| < \epsilon_n$$

on $\mathfrak{B} \cap \mathfrak{F}_{\bar{\epsilon}}$. From (37) and the estimate

$$\begin{aligned} & \|\bar{x}^n(t; t_0, x_0) - \xi^n(t; t_0, x_0)\| \\ & \leq \int_{t_0}^t \|k^{\epsilon_n}(\tau, \bar{x}^n(\tau; t_0, x_0)) - \mathfrak{K}^{\epsilon_n}(\tau, \xi^n(\tau; t_0, x_0))\| d\tau, \end{aligned}$$

where $\xi^n(\cdot; t_0, x_0)$ is a solution of (36), there follows not only the continuability to $[t_0, \bar{t}(t_0, x_0)]$ of all solutions of (36) for each n but also assertion (i) of the proposition. The final inequality of assertion (i) is a consequence of the easily demonstrated³ fact that

$$V(t_0, x_0) = \alpha(z^*(\bar{t}(t_0, x_0); t_0, x_0), G(\bar{t}(t_0, x_0))),$$

so that $t^*(t_0, x_0, u) \leq \bar{t}(t_0, x_0)$. Assertion (ii) of the proposition is of course a consequence of the definition of \mathfrak{K}^ϵ .

³ The demonstration is like that for Theorem 3(ii).

Remark 5. For obtaining approximate syntheses it is desirable that $\bar{B}_\gamma \subset \mathcal{F}_\varepsilon$. By virtue of Proposition 1, the verification of this inclusion has an interesting ramification. As a uniform limit of continuous functions (Dini's theorem and Theorem 3(ii)), $V(\cdot, \cdot)$ is continuous on \bar{B}_γ so that its set of zeros is relatively closed in $I_T \times R^n$. But the set of zeros of $V(\cdot, \cdot)$ is precisely the set of points in $I_T \times R^n$ starting from a point of which a solution of (2) can "hit" the target G .

The key element in the proof of Proposition 1, as well as in the proofs of the results of §§3 and 4, is the boundedness—uniform with respect to ϵ —of $g_p^\epsilon(\cdot, \cdot, \cdot)$. This boundedness leads directly to the uniform convergence on $[t_0, \bar{t}(t_0, x_0)]$ of a suitably selected sequence $\{\bar{x}^n(\cdot; t_0, x_0)\}$ of solutions of (35). However, it is clear that in order to have an approximate synthesis theorem it is essential that this convergence also be uniform with respect to $(t_0, x_0) \in B_\gamma$. For the establishment of this latter uniformity the boundedness of the family $\{g_p^\epsilon(\cdot, \cdot, \cdot) \mid 0 < \epsilon \leq 1\}$ does not appear to be sufficient; if, in addition, this family possesses a suitable equicontinuity property, then a synthesis theorem can be obtained. This assertion is made precise in the following proposition.

PROPOSITION 2. *In addition to the hypotheses of Proposition 1 assume that $\bar{B}_\gamma \subset \mathcal{F}_\varepsilon$ and that there exists $L > 0$ such that*

$$\|k^\epsilon(t_1, x_1) - k^\epsilon(t_2, x_2)\| \leq L\{|t_1 - t_2| + \|x_1 - x_2\|\}$$

for all ϵ , $0 < \epsilon \leq \bar{\epsilon}$, and all $(t_i, x_i) \in \bar{B}_\gamma$, $i = 1, 2$, where k^ϵ is defined by (34); then there exists a sequence $\{k^{\epsilon_n}\}$ of functions which converges on B_γ to a uniform limit function g having the following properties:

- (i) L is a uniform Lipschitz constant for g on B_γ and $g(t, x) \in \partial R(t, x)$ for all $(t, x) \in B_\gamma$;
- (ii) the solution $\bar{x}(\cdot; t_0, x_0)$ of the differential equation

$$(38) \quad \dot{x} = g(t, x), \quad x(t_0) = x_0,$$

satisfies (i), (ii) of Theorem 3;

- (iii) the uniform convergence on $[t_0, \bar{t}(t_0, x_0)]$ to $\bar{x}(\cdot; t_0, x_0)$ of the solutions of (35), with the k^{ϵ_n} taken to be those whose existence is asserted above, is also uniform with respect to $(t_0, x_0) \in B_\gamma$.

Proof. As in the proof of Proposition 1 we confine our attention to those indices n which satisfy $2^{-n} < \bar{\epsilon}$. By virtue of the Lipschitz condition, the family $\{k^{\epsilon_n}(\cdot, \cdot) \mid 2^{-n} < \bar{\epsilon}\}$ is equicontinuous on \bar{B}_γ (as well as uniformly bounded there) so that there exists a subsequence, denoted by the same indices, which possesses the asserted convergence property. That L is a uniform Lipschitz constant for the limit function g on \bar{B}_γ is the consequence of an obvious estimate.

By virtue of Theorem 4 and the continuity of $\bar{x}^n(\cdot; \cdot, \cdot)$, it follows that

for each $(t_0, x_0) \in \bar{B}_\gamma$ there exists a neighborhood η of (t_0, x_0) and a positive integer N depending on (t_0, x_0) , such that for all $(t_1, x_1) \in \eta$ and all $n > N$, $\bar{x}^n(\cdot; t_1, x_1)$ is continuable to $[t_0, \bar{i}(t_0, x_0)]$. Now \bar{B}_γ may be covered by a finite number of these η neighborhoods and so we may assert the existence of a positive integer \bar{N} , such that for all $(t_0, x_0) \in \bar{B}_\gamma$ and all $n > \bar{N}$, $\bar{x}^n(\cdot; t_0, x_0)$ is continuable to $[t_0, \bar{i}(t_0, x_0)]$. In view of this assertion and the uniqueness of solutions of (38), it follows that for each $(t_0, x_0) \in \bar{B}_\gamma$ the sequence $\{\bar{x}^n(\cdot; t_0, x_0) \mid n > \bar{N}\}$ converges uniformly on $[t_0, \bar{i}(t_0, x_0)]$ to $\bar{x}(\cdot; t_0, x_0)$ (cf. [9, Remark 2, p. 4]). The proof of (ii) may now be carried out as indicated for Theorem 3.

From (35) and (38) there is obtained the estimate

$$\begin{aligned}
 & \| \bar{x}^n(t; t_0, x_0) - \bar{x}(t; t_0, x_0) \| \\
 & \leq \int_{t_0}^t \| k^{\epsilon_n}(\tau, \bar{x}(\tau; t_0, x_0)) - g(\tau, \bar{x}(\tau; t_0, x_0)) \| d\tau \\
 (39) \quad & + \int_{t_0}^t \| k^{\epsilon_n}(\tau, \bar{x}^n(\tau; t_0, x_0)) - k^{\epsilon_n}(\tau, \bar{x}(\tau; t_0, x_0)) \| d\tau, \\
 & t_0 \leq t \leq \bar{i}(t_0, x_0).
 \end{aligned}$$

By the uniform convergence established previously, for each $\beta > 0$ there exists a positive integer Y depending only on β such that for all $n > Y$,

$$\| k^{\epsilon_n}(t, x) - g(t, x) \| < \beta$$

on \bar{B}_γ . In conjunction with (39), this bound yields

$$\begin{aligned}
 & \| \bar{x}^n(t; t_0, x_0) - \bar{x}(t; t_0, x_0) \| \\
 & \leq \beta T + \int_{t_0}^t \| k^{\epsilon_n}(\tau, \bar{x}^n(\tau; t_0, x_0)) - k^{\epsilon_n}(\tau, \bar{x}(\tau; t_0, x_0)) \| d\tau, \\
 & n > Y, \quad t_0 \leq t \leq \bar{i}(t_0, x_0);
 \end{aligned}$$

together with the Lipschitz condition satisfied by k^{ϵ_n} this last estimate yields, upon application of a well-known integral inequality, the result

$$\| \bar{x}^n(t; t_0, x_0) - \bar{x}(t; t_0, x_0) \| \leq \beta T \exp LT, \quad n > Y, \quad t_0 \leq t \leq \bar{i}(t_0, x_0).$$

This result establishes (iii) and there remains only the assertion that $g(t, x) \in \partial R(t, x)$, but this is an easy consequence of (37) and the contrary assumption.

From (34), together with the properties of W^ϵ and g_p^ϵ , one finds that k^ϵ is of class C^1 on \mathfrak{F}_ϵ so that if the conditions of Proposition 1 are satisfied, as well as $\bar{B}_\gamma \subset \mathfrak{F}_\epsilon$, then k^ϵ satisfies

$$\| k^\epsilon(t_1, x_1) - k^\epsilon(t_2, x_2) \| \leq L(\epsilon) \{ |t_1 - t_2| + \|x_1 - x_2\| \}$$

for all $(t_i, x_i) \in \bar{B}_\gamma$, $i = 1, 2$. By easy estimates one can convince oneself that there exists $L > 0$ such that $L(\epsilon) \leq L$ for $0 < \epsilon \leq \bar{\epsilon}$ —thereby satisfying the stronger Lipschitz condition of Proposition 2—*provided* the Lipschitz constants of h_p^ϵ (see the proof of Theorem 1) and W_x^ϵ do not depend on ϵ when $0 < \epsilon \leq \bar{\epsilon}$. The attempt to study the circumstances under which the latter conditions are satisfied leads once again to the type of difficulty discussed at the beginning of this section. On the one hand, there seem to be no results available concerning bounds on the partial derivatives of solutions of the Hamilton-Jacobi equation whereas, on the other hand, the complexity of the Bonnesen-Fenchel type approximants derived in Theorem 1 and [2, Theorem 4] is sufficiently severe as to prevent derivation of explicit estimates for the Lipschitz constant of h_p^ϵ .

In the light of the preceding discussion, as well as that at the beginning of this section, it appears that Theorem 4 and its corollary are about as far as one can proceed in general with the Bonnesen-Fenchel type of approximants. An alternative, which may allow one to push Hermes' approximation scheme beyond Theorem 4 and its corollary, is a family of approximants based on that discussed briefly in Remark 3. We shall define this family and summarize those of its properties which have been derived in [6, Theorem 34] and [8].

For $r_n = 2^n$ and n a sufficiently large positive integer we define⁴ $Z_n(t, x) \supset R(t, x)$ to be the closed r_n^{-1} neighborhood of $S_{r_n}(t, x)$. Then $Z_n(t, x)$ is strictly convex and smooth, where by "smooth" we mean that each point of $\partial Z_n(t, x)$ lies on a unique tangent hyperplane to $Z_n(t, x)$. Moreover, if $n_2 > n_1$ then $Z_{n_2}(t, x) \subset Z_{n_1}(t, x)$ and

$$\lim_{n \rightarrow \infty} d(Z_n(t, x), R(t, x)) = 0.$$

Finally, the continuity of $Z_n(\cdot, \cdot)$ on $I_T \times E^n$ is implied by that of $R(\cdot, \cdot)$.

These properties of Z_n imply the following properties for g^n , where $g^n(t, x, \cdot)$ is the support function of $Z_n(t, x)$: $g^n(\cdot, \cdot, \cdot)$ is continuous on $I_T \times E^n \times E^n$ and $g_p^n(\cdot, \cdot, \cdot)$ is defined and continuous on the same set. Moreover, $g_p^n(\cdot, \cdot, \cdot)$ possesses on $I_T \times E^n \times E^n$ the properties described in (b) of Theorem 1(ii). Finally, for $(t, x) \in I_T \times E^n$, the function $g_p^n(t, x, \cdot)$ is a homeomorphism of the boundary of the unit ball in E^n onto $\partial Z_n(t, x)$.

It is easy to see that Theorems 2 and 3 still hold with approximants R^ϵ replaced by approximants Z_n (cf. Remark 3), but our present state of knowledge of the Z_n does not allow a proof of the counterparts of Theorem 4 and its corollary. Inasmuch as Theorem 4 can still be proved (see [9, Exercise 9.3, p. 139]) when the C^2 requirement on the functions ψ , ω ,

⁴ By the definition of Z_n , the inclusion is proper.

g^ϵ , W^ϵ is reduced to the requirement that these functions be of class C^1 with uniformly Lipschitzian partial derivatives, it is clear that entirely satisfactory counterparts of Theorem 4 and its corollary would ensue if $g_p^n(t, x, \cdot)$ possessed a Lipschitz constant which was uniform in n . One could always obtain adequate smoothness with respect to t, x by the mollifier technique. The possession by g_p^n of such a Lipschitz constant would at the same time dispose of, to some extent, another major problem, namely that of the satisfaction of the Lipschitz hypothesis of Proposition 2. Whether g_p^n has the Lipschitz property discussed in this paragraph is still conjectural and certainly worthy of further investigation.

REFERENCES

- [1] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76-84.
- [2] H. HERMES, *The equivalence and approximation of optimal control problems*, J. Differential Equations, 1 (1965), pp. 409-426.
- [3] R. E. KALMAN, *Mathematical Optimization Techniques*, University of California Press, Berkeley, California, 1963, Chap. 16.
- [4] T. F. BRIDGLAND, JR. AND J. S. HINKEL, *The minimum miss distance problem*, Proc. Amer. Math. Soc., to appear.
- [5] T. BONNESEN AND W. FENCHEL, *Theorie der konvexen Körper*, Chelsea, New York, 1948.
- [6] H. G. EGGLESTON, *Convexity*, Cambridge University Press, London, 1958.
- [7] F. A. VALENTINE, *Convex Sets*, McGraw-Hill, New York, 1964.
- [8] T. F. BRIDGLAND, JR., *On the continuity of the r -convex hull of a continuous convex set*, to appear.
- [9] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
- [10] E. KAMKE, *Differentialgleichungen reeller Funktionen*, Akad. Verlagsgesellschaft M. B. H., Leipzig, 1930.
- [11] T. WAZEWSKI, *Sur l'appréciation du domaine d'existence des intégrales de l'équation aux dérivées partielles du premier ordre*, Ann. Polon. Math., 14 (1935), pp. 149-177.
- [12] E. KAMKE, *Differentialgleichungen, Lösungsmethoden und Lösungen. II. Partielle Differentialgleichungen erster Ordnung für eine gesuchte Funktion*, Akad. Verlagsgesellschaft, Leipzig, 1959.
- [13] V. L. DUBNOV, *Existence and uniqueness "in the large" of the solution of the Hamilton-Jacobi equation*, Vestnik Moskov. Univ. Ser. I Mat. Meh., (1965), no. 2, pp. 3-9.
- [14] T. F. BRIDGLAND, JR., *On the existence of optimal feedback controls*, this Journal, 1 (1963), pp. 261-274.

OPTIMAL SYSTEMS WITH MULTIPLE COST FUNCTIONALS*

DONG HAK CHYUNG†

1. Introduction. There are many optimal processes where an optimal control with respect to a given cost functional $C_1(u)$ is not unique. In this case it is rather natural to introduce a second cost functional $C_2(u)$ and then to choose a control which minimizes the second cost functional among the controls which minimize the first cost functional. If an optimal control chosen in this way is still not unique, then of course one can introduce a third cost functional, and so on. Then, in addition to obtaining a better optimal control, the computation of the optimal control becomes easier, for the optimal control obtained in this way is necessarily unique.

Many optimal control systems can be considered as optimal control problems with multiple cost functionals. Consider a time-optimal control system where a time-optimal control is not unique. Then one may wish to choose a control which requires minimum fuel among the time-optimal controls. A bounded phase coordinate optimal control problem can be approximated as an optimal control problem with multiple cost functionals and side constraints by introducing a penalty function and letting it be the second cost function (see [2]). It is obvious that certain differential games, too, can be considered as optimal systems with multiple cost functionals. Also the problem of steering an initial function to a given final function in an optimal control system described by a functional differential equation can be handled by the method proposed in this paper (see remarks in [1]).

As far as the existence and necessary conditions are concerned similar results can be obtained for nonlinear systems. However, for the sake of simplicity, only a linear optimal control system with two ordered cost functionals and fixed final time is studied in this paper. The results can be readily extended to systems with many cost functionals and also to systems with side constraints.

2. Problem. Consider the linear system

$$(L) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0,$$

together with

$$\begin{aligned} x^1(t) &= f^1(x(t), t) + h^1(u(t), t), & x^1(t_0) &= 0, \\ x^2(t) &= f^2(x(t), t) + h^2(u(t), t), & x^2(t_0) &= 0, \end{aligned}$$

* Received by the editors December 28, 1966, and in revised form March 15, 1967.

† Department of Electrical Engineering, University of South Carolina, Columbia, South Carolina 29208.

where $x(t) \in R^n$ is the state vector, $x^1(t) \in R^1$, $x^2(t) \in R^1$, $u(t) \in R^m$ is the control, $A(t)$ is an $n \times n$ continuous matrix, $B(t)$ is an $n \times m$ continuous matrix, $f^1(x, t) \in C^1$, $f^2(x, t) \in C^1$, $h^1(u, t) \in C^0$ and $h^2(u, t) \in C^0$ are real non-negative convex functions in $x \in R^n$ and $u \in R^m$ respectively for each t .

Let $\Omega(t_0, t_1)$ be the set of all measurable functions $u(t)$ on $[t_0, t_1]$ such that $u(t) \in \Omega$ for all t . Here $\Omega \subset R^m$ is a compact convex control constraint set. Let $C_1(u)$ and $C_2(u)$ be two cost functionals given by

$$C_1(u) = \int_{t_0}^{t_1} \{f^1(x(t), t) + h^1(u(t), t)\} dt = x^1(t_1),$$

$$C_2(u) = \int_{t_0}^{t_1} \{f^2(x(t), t) + h^2(u(t), t)\} dt = x^2(t_1).$$

Also let $G \subset R^n$ be a given convex closed target set. Denote the response of the system (L) corresponding to a control $u(t) \in \Omega(t_0, t_1)$ by $x_u(t_1)$. Then the problem is to find an optimal control $u^*(t)$ with response $x^*(t)$ from $\Omega(t_0, t_1)$ so that for a given fixed $t_1 \geq t_0$,

- (i) $x^*(t_1) \in G$,
- (ii) $C_1(u^*) \leq C_1(u)$ for all $u(t) \in \Omega(t_0, t_1)$ such that $x_u(t_1) \in G$,
- (iii) $C_2(u^*) \leq C_2(u)$ for all $u(t) \in \Omega(t_0, t_1)$ such that $x_u(t_1) \in G$ and $C_1(u) = C_1(u^*)$.

3. Sets of attainability. Let $\tilde{x} = (x, x^1) \in R^{n+1}$, $\hat{x} = (x, x^1, x^2) \in R^{n+2}$, and define new systems (\tilde{L}) and (\hat{L}) by

$$\begin{aligned} (\tilde{L}) \quad \dot{x}(t) &= A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \\ \dot{x}^1(t) &= f^1(x(t), t) + h^1(u(t), t), \quad x^1(t_0) = 0, \end{aligned}$$

and

$$\begin{aligned} (\hat{L}) \quad \dot{x}(t) &= A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \\ \dot{x}^1(t) &= f^1(x(t), t) + h^1(u(t), t), \quad x^1(t_0) = 0, \\ \dot{x}^2(t) &= f^2(x(t), t) + h^2(u(t), t), \quad x^2(t_0) = 0. \end{aligned}$$

Let $\tilde{x}_u(t) = (x_u(t), x_u^1(t))$ and $\hat{x}_u(t) = (x_u(t), x_u^1(t), x_u^2(t))$ be the responses of (\tilde{L}) and (\hat{L}) corresponding to a control $u(t) \in \Omega(t_0, t_1)$ respectively.

DEFINITION 1. The sets of attainability $K \subset R^n$, $\tilde{K} \subset R^{n+1}$ and $\hat{K} \subset R^{n+2}$ are the sets of endpoints $x_u(t_1)$, $\tilde{x}_u(t_1)$ and $\hat{x}_u(t_1)$ of the responses of the systems (L), (\tilde{L}) and (\hat{L}) respectively for all controls $u(t) \in \Omega(t_0, t_1)$, i.e.,

$$\begin{aligned} K &= \{x_u(t_1) \mid x_u(t_0) = x_0, u(t) \in \Omega(t_0, t_1)\}, \\ \tilde{K} &= \{\tilde{x}_u(t_0) \mid \tilde{x}_u(t_0) = \tilde{x}_0, u(t) \in \Omega(t_0, t_1)\}, \end{aligned}$$

\bar{K}_s and \hat{K}_s are in the same subspaces $x^1 \geq 0$ of R^{n+1} and $x^1 \geq 0, x^2 \geq 0$ of R^{n+2} respectively. Obviously \bar{K} and \bar{K}_s are the orthogonal projections of \hat{K} and \hat{K}_s on the (x, x^1) -space R^{n+1} , and K is the orthogonal projection of $\bar{K}, \bar{K}_s, \hat{K}$ and \hat{K}_s on the x -space R^n .

It is known that K is a compact convex set in R^n , \bar{K}_s is a closed convex set in R^{n+1} and \hat{K}_s is a closed convex set in R^{n+2} (see [3], [4]). Also, from the properties of the system (\hat{L}) it is easy to show that \hat{K} is bounded. Let $z^1 = \sup x^1$ and $z^2 = \sup x^2$ for all $(x, x^1, x^2) \in \hat{K}$, and define the set \hat{M} by

$$\hat{M} = \{\hat{x} = (x, x^1, x^2) \mid \hat{x} \in \hat{K}_s, x^1 \leq z^1, x^2 \leq z^2\}.$$

Then $\hat{K} \subset \hat{M} \cap \hat{K}_s$ and \hat{M} is compact in R^{n+2} , for $x \in K$ if $(x, x^1, x^2) \in \hat{M}$ and K is compact.

Define the target set \hat{G} in R^{n+2} by

$$\hat{G} = \{\hat{x} = (x, x^1, x^2) \mid x \in G, -\infty < x^1 < \infty, -\infty < x^2 < \infty\},$$

that is, $\hat{G} = G \times R^1 \times R^1$. Clearly \hat{G} is closed and convex in R^{n+2} , for G is closed and convex in R^n . Then the optimal control problem is to find an optimal point $\hat{x}^* = (x^*, x^{1*}, x^{2*})$ in $\hat{K} \cap \hat{G}$ such that $x^{1*} = \min x^1$ for all $(x, x^1, x^2) \in \hat{K} \cap \hat{G}$ and $x^{2*} = \min x^2$ for all $(x, x^1, x^2) \in \hat{K} \cap \hat{G}$. The control which steers the corresponding response to the optimal point \hat{x}^* is an optimal control.

4. Existence.

THEOREM 1. *Suppose there exists a control $u(t) \in \Omega(t_0, t_1)$ which steers the response $x(t)$ to G at $t = t_1$. Then there exists an optimal control.*

Proof. By the assumption $K \cap G \neq \emptyset$. Then $\bar{K}_s \cap \hat{M} \cap \hat{G} \neq \emptyset$ for $\hat{K} \subset \bar{K}_s \cap \hat{M}$. On the other hand, $\bar{K}_s \cap \hat{M} \cap \hat{G}$ is compact, for \bar{K}_s and \hat{G} are closed and \hat{M} is compact. Therefore there exists a point $\hat{y} = (y, y^1, y^2)$ in $\bar{K}_s \cap \hat{M} \cap \hat{G}$ with $y^1 = \min x^1$ for all $\hat{x} = (x, x^1, x^2)$ in $\bar{K}_s \cap \hat{M} \cap \hat{G}$. Let P be the hyperplane defined by $x^1 = y^1$ in R^{n+2} . Then $\bar{K}_s \cap \hat{M} \cap \hat{G} \cap P$ is again a nonempty compact set and so there exists a point $\hat{x}^* = (x^*, x^{1*}, x^{2*})$ in $\bar{K}_s \cap \hat{M} \cap \hat{G} \cap P$ such that $x^{2*} = \min x^2$ for all $\hat{x} = (x, x^1, x^2)$ in $\bar{K}_s \cap \hat{M} \cap \hat{G} \cap P$. Note that $x^{1*} = y^1$. Since $\hat{G} = G \times R^1 \times R^1$, it is obvious that $x^{1*} = \min x^1$ and $x^{2*} = \min x^2$ for all $\hat{x} = (x^0, x, x^{n+1})$ in \bar{K}_s with $x = x^*$. Therefore, from the definition of the saturation set, \hat{x}^* must be a point of \hat{K} and so $\hat{x}^* \in \hat{K} \cap \hat{G}$. Furthermore, since it is an extremal point, \hat{x}^* must be a boundary point of both \hat{K} and \bar{K}_s , that is, $\hat{x}^* \in \partial \hat{K} \cap \partial \bar{K}_s$. It is easy to show, from the definition of the point \hat{x}^* , $x^{1*} = \min x^1$ for all (x, x^1, x^2) in $\hat{K} \cap \hat{G}$ and $x^{2*} = \min x^2$ for all (x, x^{1*}, x^2) in $\hat{K} \cap \hat{G}$, for $\hat{K} \subset \bar{K}_s \cap \hat{M}$. Hence \hat{x}^* is an optimal point and there exists an optimal control.

5. Necessary and sufficient conditions (maximum principle).

DEFINITION 4. A control $u(t) \in \Omega(t_0, t_1)$ with response $x(t)$ is called *maximal* if there exists a nontrivial response $\hat{p}(t) = (p(t), p^1, p^2) \in \mathbb{R}^{n+2}$ of the equation

$$\begin{aligned} p^1 &\leq 0, \text{ constant,} \\ p^2 &\leq 0, \text{ constant,} \\ \dot{p}(t) &= -p(t)A(t) - p^1 \frac{\partial f^1}{\partial x}(x(t), t) - p^2 \frac{\partial f^2}{\partial x}(x(t), t), \end{aligned}$$

such that

$$\begin{aligned} p^1 h^1(u(t), t) + p^2 h^2(u(t), t) + p(t)B(t)u(t) \\ = \max_{u \in \Omega} \{p^1 h^1(u, t) + p^2 h^2(u, t) + p(t)B(t)u\} \end{aligned}$$

almost everywhere on $[t_0, t_1]$.

THEOREM 2. A control $u(t) \in \Omega(t_0, t_1)$ steers the corresponding response $\hat{x}_u(t)$ to the common boundary $\partial \hat{K} \cap \partial \hat{K}_s$ at $t = t_1$ if and only if $u(t)$ is a maximal control.

We omit the proof, for a proof can be found in [3]. It was also shown in the proof that $\hat{p}(t_1)$ is an exterior normal vector to \hat{K}_s at $\hat{x}_u(t_1) \in \partial \hat{K}_s$, i.e., $\hat{p}(t_1)$ is normal to the supporting hyperplane π of \hat{K}_s at $\hat{x}_u(t_1)$ directing into the half-space defined by π which does not contain \hat{K}_s . Such a supporting hyperplane π always exists because \hat{K}_s is convex and $\hat{x}_u(t_1) \in \partial \hat{K}_s$.

THEOREM 3. If there exists a control in $\Omega(t_0, t_1)$ which steers the response of the system (L) to G at $t = t_1$, then there exists a nontrivial solution $x^*(t)$, $q^*(t)$, $p^*(t)$ of

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u^*(t, q, p), \\ \dot{q}(t) &= -q(t)A(t) - p^1 \frac{\partial f^1}{\partial x}(x(t), t), \\ \dot{p}(t) &= -p(t)A(t) - p^1 \frac{\partial f^1}{\partial x}(x(t), t) - p^2 \frac{\partial f^2}{\partial x}(x(t), t), \end{aligned}$$

with $x(t_0) = x_0$, and either

(a) $x(t_1) \in G$, $p^1 < 0$, $p^2 < 0$, $q(t_1) = p(t_1) = 0$, or

(b) $x(t_1) \in \partial G$, $p^1 \leq 0$, $p^2 \leq 0$, $q(t_1) = p(t_1)$ interior normal to G at $x(t_1)$.

Here $u^*(t, q, p)$ is defined by the maximum principle

$$\begin{aligned} p^1 h^1(u^*, t) + q(t)B(t)u^* &= \max_{u \in \Omega} \{p^1 h^1(u, t) + q(t)B(t)u\}, \\ p^1 h^1(u^*, t) + p^2 h^2(u^*, t) + p(t)B(t)u^* &= \max_{u \in \Omega} \{p^1 h^1(u, t) + p^2 h^2(u, t) \\ &\quad + p(t)B(t)u\}. \end{aligned}$$

If $p^1 \neq 0$ and $p^2 \neq 0$, then $u^*(t) = u^*(t, q^*, p^*)$ with response $x^*(t)$, $q^*(t)$, $p^*(t)$ is an optimal control.

Proof. By Theorem 1 there exists an optimal point \hat{x}^* in \hat{K} , and $\hat{x}^* \in \partial\hat{K} \cap \partial\hat{K}_s$. Therefore, by Theorem 2 there exists a solution $\hat{p}^*(t)$, and the control $u^*(t) \in \Omega(t_0, t_1)$, which steers the response to \hat{x}^* at $t = t_1$, satisfies the second maximum condition. Furthermore $\hat{p}^*(t_1)$ is an exterior normal vector to the convex set \hat{K}_s at $\hat{x}^* \in \partial\hat{K}_s$ and so $p^{1*} \leq 0$, $p^{2*} \leq 0$. The first maximum condition together with the existence of $q^*(t)$ follows from the fact that $u^*(t)$ minimizes the first cost functional $C_1(u)$ (see [4] for a proof).

Conversely, if a control $u^*(t)$ satisfies the first maximum condition with the given endpoint conditions, then $u^*(t)$ minimizes the first cost functional $C_1(u) = x^1(t_1)$ (again see [4] for a proof). Now, in addition to this, if $u^*(t)$ satisfies the second maximum condition, then the corresponding response $\hat{x}^*(t_1)$ is on $\partial\hat{K} \cap \partial\hat{K}_s$ and $\hat{p}^*(t_1)$ is an exterior normal vector to \hat{K}_s at $\hat{x}^*(t_1)$. If the endpoint conditions either (a) or (b) are satisfied, then from Fig. 1 it is apparent that $\hat{x}^*(t_1) = (x^*(t_1), x^{1*}(t_1), x^{2*}(t_1))$ is a point with $x^{2*}(t_1) = \min x^2$ for all $\hat{x} = (x, x^1, x^2)$ in \hat{K} such that $x \in G$, $x^1 = x^{1*}(t_1)$. But then since $x^{1*}(t_1) = C_1(u^*)$ is minimum for all \hat{x} in $\hat{K} \cap \hat{G}$, $\hat{x}^*(t_1)$ is an optimal point and so $u^*(t)$ is an optimal control.

Remark. If either p^1 or p^2 is zero, then an optimal control must be given by $u^*(t) = u^*(t, q^*, p^*)$. However this is no longer sufficient to be an optimal control.

6. Remarks. The above results can easily be extended to the case when the first cost functional $C_1(u)$ is given by

$$C_1(u) = g(x(t_1)) + \int_{t_0}^{t_1} \{f^1(x(t), t) + h^1(u(t), t)\} dt,$$

where $g(x) \in C^1$ is a convex function in $x \in R^n$. Furthermore, if $G = R^n$ (the free endpoint problem), then Theorem 4 is valid as it is except that the endpoint conditions should be replaced by the following simpler conditions:

$$q(t_1) = p(t_1) = -\text{grad } g(x(t_1)), \quad p^1 < 0, p^2 < 0.$$

Also similar results can be obtained when $\Omega = R^m$. In this case it should be assumed that $h^1(u, t) \geq a |u|^k$, $h^2(u, t) = a |u|/k$ for some $a > 0$, $k > 1$ for all $u \in R^m$ to avoid impulse type controls.

REFERENCES

- [1] D. H. CHYUNG AND E. B. LEE, *On certain extremal problems involving functional differential equation models*, Proceedings of Conference on the Mathematical Theory of Control, Academic Press, New York, 1967.
- [2] E. B. LEE, *An approximation to linear bounded phase coordinate control problem*, J. Math. Anal. Appl., 13 (1966), pp. 550-564.

- [3] ———, *Linear optimal control problems with isoperimetric constraints*, IEEE Trans. Automatic Control, AC-12 (1967), pp. 87–90.
- [4] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [5] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [6] L. A. ZADEH, *Optimality and non-scalar-valued performance criteria*, IEEE Trans. Automatic Control, AC-8 (1963), pp. 59–60.

NECESSARY OPTIMALITY CONDITIONS FOR DISTRIBUTED-PARAMETER SYSTEMS*

A. I. EGOROV†

The mathematical theory of optimal control processes, created by L. S. Pontryagin and his coworkers, allows us to find an optimal control in the case when the process is described by ordinary differential or difference-differential equations [1]. The maximum principle is the basic mathematical tool in this theory and at first it appeared as a necessary condition for optimality. Later it was shown [2] that this condition also is sufficient for a rather wide group of problems.

In a number of important practical cases the control law of the systems cannot be given by ordinary differential or difference-differential equations but can be described by partial differential equations (for example, see [3]–[6]). Control systems of such type are called distributed-parameter systems.

In this article we shall consider certain optimal control problems in processes which can be described by the classical boundary value problems for hyperbolic and parabolic equations. In general, the ranges of the independent variables are not fixed.

The article consists of two parts. In Part I the necessary optimal conditions are obtained in the case when the process is described by hyperbolic equations with Goursat boundary conditions. The method presented in this part can be applied also in the case when the Goursat conditions are replaced by conditions of other types. In Part II the control problem is considered in the case when the process can be described by boundary value problems for second-order parabolic equations. A different method for solving certain such problems has been presented by Yu. V. Egorov in [7].

I. OPTIMAL PROCESSES IN SYSTEMS WHOSE BEHAVIOR IS DESCRIBED BY HYPERBOLIC EQUATIONS

1.1. Statement of the problem. Optimality conditions. We consider a system in which control is effected by a law given by the differential equations

$$(1.1) \quad \begin{aligned} z_{ixy} &= f_i(x, y, z_1, \dots, z_n, z_{1x}, \dots, z_{nx}, z_{1y}, \dots, z_{ny}, v), \\ i &= 1, \dots, n. \end{aligned}$$

* Originally published in *Mat. Sb.*, 69(111) (1966), pp. 371–421. Submitted for publication November 25, 1964. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid from the National Science Foundation.

† Institute of Automation, Academy of Sciences of the Khirgiz SSR, Frunze.

We assume that the functions $f_i(x, y, z, u, w, v)$ are defined, are continuously differentiable in x and y for $0 \leq x \leq X$, $0 \leq y \leq Y$, and are twice-continuously differentiable in all the remaining arguments in some open region A of the variables z, u, v, w . The parameter v can take values in some open or closed region V of an r -dimensional Euclidean space.

The Goursat boundary conditions

$$(1.2) \quad z_i(0, y) = \phi_i^1(y), \quad z_i(x, 0) = \phi_i^2(x), \quad i = 1, \dots, n,$$

are imposed on the functions z_i given by (1.1); the functions $\phi_i^1(y)$ and $\phi_i^2(x)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, are continuous, piecewise-continuously-differentiable and satisfy the conjugacy relations $\phi_i^1(0) = \phi_i^2(0)$.

As controls we shall take piecewise-continuous functions $v = v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, taking values in the region V . We assume that the lines of discontinuity of these functions are piecewise-smooth. We do not exclude the possibility here that the form of the dependency of certain specified components of the vector-valued function $v(x, y)$ on the arguments x and y may be given in advance. For example, the individual components may be functions of only one variable (x or y), may depend only on $x \pm y$, etc.

If the control $v(x, y)$ is continuous, to it corresponds a unique solution $z(x, y) = \{z_1(x, y), \dots, z_n(x, y)\}$ of the Goursat problem (1.1)–(1.2); moreover, the functions $z_i(x, y)$ have the continuous derivatives z_{ix} , z_{iy} and z_{ixy} .

However, if $v(x, y)$ has lines of discontinuity Γ , then in order to determine uniquely the corresponding solution of problem (1.1)–(1.2), smoothness conditions for the functions $z_i(x, y)$ on Γ should be given in advance (for example, see [8]). It turns out here that, in general, the solution has piecewise-continuous derivatives z_{ix} , z_{iy} and z_{ixy} . Therefore, in what follows we shall consider, just as we did in [9], that with every control there is associated a class of functions in which the problem (1.1)–(1.2) has a unique solution for the control selected.

From the set of controls we pick out the class of admissible controls consisting of those functions $v(x, y)$ for which the corresponding solutions of the Goursat problem (1.1)–(1.2) lie in a region where the functions f_i satisfy the conditions indicated above.

We shall say that an admissible control $v(x, y)$ transfers the system from the state (1.2) to the state (1.3) if the corresponding solution of problem (1.1)–(1.2) satisfies the conditions

$$(1.3_1) \quad F_\alpha(X, Y, z(X, Y)) = 0, \quad \alpha = 1, \dots, j,$$

$$(1.3_2) \quad \psi_\beta^1(X, y, z(X, y)) = 0, \quad \beta = 1, \dots, k,$$

$$\psi_\gamma^2(x, Y, z(x, Y)) = 0, \quad \gamma = 1, \dots, l,$$

$$(1.3_3) \quad \int_0^x \Phi_\delta(x, Y, z(x, Y), z_x(x, Y)) dx - a_\delta = 0, \quad \delta = 1, \dots, m,$$

$$(1.3_4) \quad \int_0^Y \Psi_\epsilon(X, y, z(X, y), z_y(X, y)) dy - b_\epsilon = 0, \quad \epsilon = 1, \dots, \nu,$$

where $j + k + l + m + \nu \leq n$, a_δ and b_ϵ are some constants. The functions F_α , ψ_β , Φ_δ , Ψ_ϵ satisfy exactly the same conditions as the functions f_i do and, moreover, $\psi_\beta^1(X, 0, \phi^2(X)) = \psi_\gamma^2(0, Y, \phi^1(Y)) = 0$. In general, the quantities X and Y are not fixed in advance and may vary from one admissible control to another.

We pose the problem: from among the admissible controls transferring the system from the state (1.2) to the state (1.3), find that control for which the functional

$$(1.4) \quad S = \sum_{i=1}^n A_i z_i(X, Y),$$

where the A_i are real constants, takes the smallest possible value.

The control and the corresponding solution of problem (1.1)–(1.2), which together constitute the solution of the problem we have posed, will be called, respectively, the *control optimal relative to S* and the *optimal solution*.

For example, let $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transfer the system from the state (1.2) to the state (1.3), and let $z(x, y)$ be the corresponding solution of problem (1.1)–(1.2). In general, it is impossible to assert that another control, different from v , transferring the system from the state (1.2) to the state (1.3), exists in the class of admissible controls. If there are no such controls or if there is a finite number of them, the stated optimal problem becomes a nonvariational one. It is necessary to seek out all the controls which transfer the system from the one state to the other. Therefore, we shall assume a priori that the class of admissible controls is sufficiently complete.

The results set forth below are obtained on the assumption that the completeness of the class of admissible controls is defined by the following basic properties.

Let $v(x, y)$ and $v_1(x, y)$ be some admissible controls defined in the regions D , $0 \leq x \leq X$, $0 \leq y \leq Y$, and D_1 , $0 \leq x \leq X_1$, $0 \leq y \leq Y_1$, and transferring the system from the state (1.2) to the state (1.3). Then, for an arbitrarily small positive number ϵ we can find a control $v_\epsilon(x, y)$ defined in the region D_ϵ , $0 \leq x \leq X_\epsilon$, $0 \leq y \leq Y_\epsilon$, and such that:

- (i) it transfers the system from the state (1.2) to the state (1.3) and, moreover, (1.3) is satisfied when $X = X_\epsilon$, $Y = Y_\epsilon$;
- (ii) it is defined in the region $D_\epsilon \times D$ by the formula

$$v_{\epsilon}(x, y) = \begin{cases} v_1(x, y) & \text{when } (x, y) \in G_{\epsilon}, \\ v(x, y) & \text{when } (x, y) \in D_{\epsilon} \times D \setminus G_{\epsilon}, \end{cases}$$

where G_{ϵ} is an arbitrary given region (whose area equals ϵ) lying strictly within D ;

(iii) the inequalities $|X - X_{\epsilon}| \leq L\epsilon$ and $|Y - Y_{\epsilon}| \leq L\epsilon$ are satisfied, where the constant L is independent of ϵ .

This completeness condition is not investigated in this paper. It is possible that it is a corollary of other simpler conditions which are capable of being verified directly. In classical variational problems it is replaced by the property which in essence is that every admissible surface (curve) is contained in a family of surfaces (curves) of the same kind, depending on one or several parameters (for example, see [10, p. 277]).

We can reduce to the problem being considered analogous optimal control problems in which is chosen, as the optimality criterion, a functional of one of the following types:

$$\int_0^x \int_0^Y f_0(x, y, z, z_x, z_y, v) dy dx, \quad \int_0^Y f_0(X, y, z(X, y), z_y(X, y)) dy,$$

$$\int_0^x f_0(x, Y, z(x, Y), z_x(x, Y)) dx, \quad f_0(X, Y, z(X, Y)),$$

where f_0 is a function of the same type as the f_i in (1.1). To do this it suffices to introduce an auxiliary function $z_0(x, y)$, as was done in [9]. Such an operation makes no change whatsoever in conditions (1.3).

The introduction of the supplementary conditions (1.3) into the optimal problem allows us to study a sufficiently wide group of problems which are of theoretical and practical interest. For example, if in these conditions only the functions F_{α} , $\alpha = 1, \dots, j \leq n$, are nonzero, this signifies that the point $z(X, Y)$ determined by the solution of problem (1.1)–(1.2) must lie in a certain manifold in the space of the variables z_1, \dots, z_n , which in particular, may degenerate to a point. In this latter case conditions (1.3) can be written as

$$(1.3') \quad z_i(X, Y) - z_i^1 = 0, \quad i = 1, \dots, n,$$

where z_1^1, \dots, z_n^1 are given numbers. If only the relation (1.3₂) occurs in the conditions (1.3), then the functions $z_i(x, Y)$ and $z_i(X, y)$ should belong to a certain manifold. In particular, these conditions may signify that

$$(1.3'') \quad z_i(X, y) = \psi_i^1(y), \quad z_i(x, Y) = \psi_i^2(x), \quad i = 1, \dots, n,$$

where $\psi_i^1(y)$ and $\psi_i^2(x)$ are given functions. Thus, in this case we are required to transfer the system from one state to another in optimal fashion,

where, moreover, each of these states is given by $2n$ functions which determine the "initial" and "terminal" values of the functions $z_i(x, y)$.

The time-optimal problem also can be considered as a special case of the problem we have formulated. Indeed, in a number of physical problems the solution of which reduces to the investigation of (1.1) with supplementary conditions (1.2), one of the independent variables is time (we denote it by y) while the other, x , is a space variable indicating position within the plant. Therefore, in such problems the quantity X can be taken as a given constant. If as the optimality criterion we take the functional $S_T = Y$ and we adopt the supplementary conditions in the form (1.3''), we obtain the time-optimal problem in which the terminal state of the system is given by the functions $\psi_i^1(y)$ and $\psi_i^2(x)$ defined on the lines $x = X$ and $y = Y$. If, however, the supplementary conditions are given in the form

$$\int_0^X \Phi_i(x, z(x, Y), z_x(x, Y)) dx = a_i, \quad i = 1, \dots, n,$$

we obtain another time-optimal problem in which the terminal state of the system is given in an integral form. In order to reduce such problems to the problem with an optimality criterion of form (1.4), we introduce a new variable $z_0(x, y)$ by setting

$$(1.5) \quad z_{0xy} = 0, \quad z_0(x, 0) = 0, \quad z_0(0, y) = y, \quad 0 \leq x \leq X, 0 \leq y \leq Y.$$

Then the functional S_T can be written as $S_T = z_0(X, Y)$, but the supplementary conditions remain unaltered.

In order to formulate the necessary conditions for the optimality of the control in the stated optimization problem, we take an arbitrary admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, and we denote by $z(x, y)$, the corresponding solution of the system (1.1) with conditions (1.2). With these functions we associate the auxiliary functions $u_i(x, y)$, $i = 1, \dots, n$, $C_\beta(y)$, $\beta = 1, \dots, k$, and $B_\gamma(x)$, $\gamma = 1, \dots, l$, and also the constants p_α , $\alpha = 1, \dots, j$, q_δ , $\delta = 1, \dots, m$, and r_ϵ , $\epsilon = 1, \dots, \nu$ ($j + k + l + m + \nu \leq n$, see conditions (1.3)), which we determine from the relations (here and in the following the symbols d/dx and d/dy denote the total derivatives with respect to x and y , respectively):

$$(1.6) \quad u_{ixy} = \frac{\partial H}{\partial z_i} - \frac{d}{dx} \left(\frac{\partial H}{\partial z_{ix}} \right) - \frac{d}{dy} \left(\frac{\partial H}{\partial z_{iy}} \right),$$

$$0 \leq x \leq X, \quad 0 \leq y \leq Y,$$

$$(1.7) \quad u_{ix}(x, Y) + \frac{\partial H}{\partial z_{iy}} = - \sum_{\gamma=1}^l \frac{\partial \psi_\gamma^2}{\partial z_i} B_\gamma(x) - \sum_{\delta=1}^m q_\delta \left[\frac{\partial \Phi_\delta}{\partial z_i} - \frac{d}{dx} \left(\frac{\partial \Phi_\delta}{\partial z_{ix}} \right) \right] \quad \text{when } y = Y,$$

$$(1.8) \quad u_{iy}(X, y) + \frac{\partial H}{\partial z_{ix}} = - \sum_{\beta=1}^k \frac{\partial \psi_{\beta}^1}{\partial z_i} C_{\beta}(y) \\ - \sum_{\epsilon=1}^p r_{\epsilon} \left[\frac{\partial \Psi_{\epsilon}}{\partial z_i} - \frac{d}{dy} \left(\frac{\partial \Psi_{\epsilon}}{\partial z_{iy}} \right) \right] \quad \text{when } x = X,$$

$$(1.9) \quad u_i(X, Y) = - \left[A_i + \sum_{\alpha=1}^j p_{\alpha} \frac{\partial F_{\alpha}}{\partial z_i} + \sum_{\epsilon=1}^p r_{\epsilon} \frac{\partial \Psi_{\epsilon}}{\partial z_{iy}} \sum_{\delta=1}^m q_{\delta} \frac{\partial \Phi_{\delta}}{\partial z_{ix}} \right] \\ \text{when } x = X, y = Y,$$

$$(1.10) \quad \sum_{i=1}^n A_i z_{ix}(X, Y) + \sum_{\alpha=1}^j p_{\alpha} \frac{dF_{\alpha}(X, Y, z(X, Y))}{dX} \\ + \sum_{\delta=1}^m q_{\delta} \Phi_{\delta}(X, Y, z(X, Y), z_x(X, Y)) \\ + \sum_{\epsilon=1}^p r_{\epsilon} \int_0^Y \frac{d\Psi_{\epsilon}(X, y, z(X, y), z_y(X, y))}{dX} dy \\ + \sum_{\beta=1}^k \int_0^Y C_{\beta}(y) \frac{d\psi_{\beta}^1(X, y, z(X, y))}{dX} dy = 0,$$

$$(1.11) \quad \sum_{i=1}^n A_i z_{iy}(X, Y) + \sum_{\alpha=1}^j p_{\alpha} \frac{dF_{\alpha}(X, Y, z(X, Y))}{dY} \\ + \sum_{\delta=1}^m q_{\delta} \int_0^X \frac{d\Phi_{\delta}(x, Y, z(x, Y), z_x(x, Y))}{dY} dx \\ + \sum_{\epsilon=1}^p r_{\epsilon} \Psi_{\epsilon}(X, Y, z(X, Y), z_y(X, Y)) \\ + \sum_{\gamma=1}^l \int_0^X B_{\gamma}(x) \frac{d\Psi_{\gamma}^2(x, Y, z(x, Y))}{dY} dx = 0,$$

where we have introduced the notation

$$(1.12) \quad H(x, y, U, v) = \sum_{i=1}^n u_i f_i(x, y, z, z_x, z_y, v), \\ U = (z_1, \dots, z_n, z_{1x}, \dots, z_{nx}, z_{1y}, \dots, z_{ny}, u_1, \dots, u_n).$$

In the general case the functions z_{ixx} , z_{iyj} , v_x and v_y occur in the right-hand sides of (1.6). However, because of the abovementioned conditions imposed on the functions f_i and on the admissible controls, these derivatives should not exist. Therefore, in what follows we shall assume that the functions f_i are representable in the following form

$$f_i = \sum_{j,k=1}^n a_{ijk}(x, y, z) z_{jx} z_{ky} + \sum_{j=1}^n b_{ij}(x, y, z) z_{jx} \\ + \sum_{k=1}^n c_{ik}(x, y, z) z_{ky} + d_i(x, y, z, v),$$

where the functions a_{ijk} , b_{ij} , c_{ik} and d_i are continuously differentiable in x and y and twice continuously differentiable in all the remaining arguments. If it happens that a_{ijk} , b_{ij} , c_{ik} depend only on v , then we should require of the admissible controls that they have piecewise-continuous derivatives v_x and v_y . When these conditions are fulfilled we can consider that to every collection of numbers p_α , q_δ and r_ϵ and also to the functions $B_\gamma(x)$ and $C_\beta(y)$ of the system of relations (1.6)–(1.9), we can associate a unique vector $u(x, y)$ defined in the region $0 \leq x \leq X$, $0 \leq y \leq Y$.

Indeed, let the numbers p_α , q_δ , r_ϵ and the functions $B_\gamma(x)$, $C_\beta(y)$ be given. Then we can find the quantities $u_i(X, Y)$ uniquely from conditions (1.9). Consequently, for the determination of $u_i(x, Y)$ and $u_i(X, y)$ we obtain a system of ordinary differential equations (1.8) and (1.9) with the initial conditions

$$u_i(x, Y) |_{x=X} = u_i(X, Y), \quad u_i(X, y) |_{y=Y} = u_i(X, Y).$$

By virtue of the conditions satisfied by the functions f_i , the functions $u_i(x, Y)$ and $u_i(X, y)$ are uniquely determined. Thus, for the determination of the functions $u_i(x, y)$, we obtain the system of equations (1.6) with the Goursat boundary conditions

$$u_i(x, y) |_{x=X} = u_i(X, y), \quad u_i(x, y) |_{y=Y} = u_i(x, Y).$$

In accordance with the conditions indicated above such a problem has a unique solution. The purpose of the supplementary conditions (1.10) and (1.11), and the assignment of the constants p_α , q_δ , r_ϵ and of the functions $B_\gamma(x)$, $C_\beta(y)$, will be indicated in the analysis of the optimality conditions formulated below as Theorem 1.

We shall take the functional

$$(1.13) \quad J[v] = \int_0^X \int_0^Y H(x, y, U(x, y), v) dy dx,$$

defined on the admissible controls.

We shall say that an admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transferring the system from the state (1.2) to the state (1.3), satisfies a maximum condition relative to the function $u(x, y)$ if for any other admissible control $v_1(x, y)$, $0 \leq x \leq X_1$, $0 \leq y \leq Y_1$, the inequality

$$(1.14) \quad \Delta J[v] = \iint_G [H(x, y, U(x, y), v_1) - H(x, y, U(x, y), v)] dy dx \leq 0$$

is satisfied, where the vector $U(x, y) = (z, z_x, z_y, u)$ corresponds to the control $v(x, y)$, i.e., in this vector the function $z(x, y)$ is determined as the solution of problem (1.1)–(1.2) corresponding to this control. Let G denote the intersection of the regions D , $0 \leq x \leq X$, $0 \leq y \leq Y$, and D_1 , $0 \leq x \leq X_1$, $0 \leq y \leq Y_1$.

The following theorem gives the necessary optimality conditions.

THEOREM 1. *In order that an admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transferring the system from the state (1.2) to the state (1.3), and the corresponding solution $z(x, y)$ of problem (1.1)–(1.2) be optimal relative to functional (1.4), it is necessary that there exist functions $u(x, y)$, $B(x)$ and $C(y)$, and also constants p_α , q_δ , r_ϵ such that:*

(i) *the collection of functions $z(x, y)$, $u(x, y)$, $B(x)$, $C(y)$, $v(x, y)$ and the constants p_α , q_δ , r_ϵ satisfy (1.1) and (1.6) and also the conditions (1.2) and (1.7)–(1.11);*

(ii) *the control $v(x, y)$ satisfies the maximum condition relative to the function $u(x, y)$.*

Before we proceed to prove this theorem we analyze it and obtain some corollaries for various special cases which are of specific practical interest.

1.2. Discussion of Theorem 1 and some of its corollaries. Let us show that Theorem 1 yields a “complete” system of conditions for picking out the controls among which there may be optimal ones.

Indeed, the maximum condition (1.4) can be interpreted as an equation from which v is determined as a function (possibly, nonunique) of the vector $U: v = v(U)$. Substituting it into (1.1) and (1.6) we get a system of $2n$ second order equations in the $2n$ functions z_i and u_i . Its general solution depends on $4n$ arbitrary functions. To eliminate $2n$ of these functions we use the boundary conditions (1.2). To eliminate the remaining $2n$ arbitrary functions we use conditions (1.7)–(1.9). Conditions (1.7) and (1.8) are ordinary differential equations in the variables $u_i(x, Y)$ and $u_i(X, y)$. The general solution of these equations depends on $2n$ arbitrary constants and on the parameters p_α , q_δ , r_ϵ , X , Y , and also on the undetermined functions $B_\gamma(x)$ and $C_\beta(y)$. The arbitrary constants are eliminated with the aid of conditions (1.9). Thus, the functions $u_i(x, Y)$ and $u_i(X, y)$ remain dependent on the parameters p_α , q_δ , r_ϵ , X , Y and on the undetermined functions $B_\gamma(x)$ and $C_\beta(y)$. The $2n$ arbitrary functions remaining as a result of solving (1.1) and (1.6) with boundary conditions (1.2) are eliminated with the aid of the conditions

$$u_i(x, y) \big|_{x=X} = u_i(X, y), \quad u_i(x, y) \big|_{y=Y} = u_i(x, Y).$$

As a result of such elimination there now remain the undetermined parameters p_α , q_δ , r_ϵ , X , Y and functions $B_\gamma(x)$ and $C_\beta(y)$. To eliminate these we have the relations (1.3), (1.10) and (1.11) which in number equal the number of undetermined quantities. As a result we obtain a “complete” system of relations for the elimination of all undetermined quantities. Having computed the vector $U = (z, z_x, z_y, u)$ in this manner and having substituted it in the function $v = v(U)$ found from (1.14), we obtain all the controls which satisfy the conditions of the theorem. If it turns out that

there is a finite number of such controls, and if from the physics of the problem an optimal control should exist, then it can be found by testing in sequence the controls picked out by the theorem. From these discussions it follows that, although Theorem 1 does not give sufficient optimality conditions, it can be utilized in the solution of practical problems.

When defining the set of controls we remarked that the form of the dependence of individual components of the vector $v(x, y)$ on the arguments x and y may be given in advance. For example, certain components of vector v may depend on only one variable (x or y).

If this requirement is removed, i.e., if as admissible controls we take all piecewise-continuous functions $v(x, y)$ with values in the control region V , then inequality (1.14) is equivalent to the conditions

$$(1.14') \quad H(x, y, U(x, y), v(x, y)) ((=)) \sup_{v \in V} H(x, y, U(x, y), v),$$

where the symbol $((=))$ denotes equality valid for all x and y of the region G except for points (x, y) lying on a finite number of lines with zero area. We can verify the validity of this assertion by contradiction.

Indeed, let us first assume that condition (1.14') does not follow from condition (1.14). Then we can find a control $\tilde{v}(x, y)$ defined in the region \tilde{G} , $0 \leq x \leq \tilde{X}$, $0 \leq y \leq \tilde{Y}$, such that, first, the inequality

$$\iint_{D_2} [H(x, y, U(x, y), \tilde{v}(x, y)) - H(x, y, U(x, y), v(x, y))] dy dx \leq 0$$

is valid, where $D_2 = D \times \tilde{G}$, and, second, there is a point (\tilde{x}, \tilde{y}) in D_2 at which

$$(1.15) \quad H(\tilde{x}, \tilde{y}, U(\tilde{x}, \tilde{y}), \tilde{v}(\tilde{x}, \tilde{y})) > H(\tilde{x}, \tilde{y}, U(\tilde{x}, \tilde{y}), v(\tilde{x}, \tilde{y})).$$

Since the function $v(x, y)$ is piecewise-continuous by definition, the functions $u(x, y)$ and $z(x, y)$ corresponding to this control are continuous, but z_x and z_y may possess discontinuities of the first kind. Therefore, we can find a region D_ϵ belonging to D_2 and having an area ϵ such that it contains the point (\tilde{x}, \tilde{y}) and such that the inequality

$$H(x, y, U(x, y), v(x, y)) < H(x, y, U(x, y), \tilde{v}(x, y))$$

is satisfied for all $(x, y) \in D_\epsilon$. We construct an auxiliary control $v_1(x, y)$, defined in the region D_3 , $0 \leq x \leq X_3$, $0 \leq y \leq Y_3$, which, in general, can include D within itself. Of this control we require that:

- (i) it transfer the system from the state (1.2) to the state (1.3);
- (ii) it be defined in the region $D_3 \times D$ by the formula

$$v_1(x, y) = \begin{cases} \tilde{v}(x, y) & \text{if } (x, y) \in D_\epsilon, \\ v(x, y) & \text{if } (x, y) \in D_3 \times D \setminus D_\epsilon; \end{cases}$$

(iii) it be defined arbitrarily at the points (x, y) not belonging to the region D while remaining admissible and satisfying condition (i).

Then, by virtue of inequality (1.15),

$$\iint_{D \times D} [H(x, y, U(x, y), v_1(x, y)) - H(x, y, U(x, y), v(x, y))] dy dx > 0,$$

but this contradicts condition (1.14).

Analogously we can show that condition (1.14) follows from (1.14').

Theorem 1 gives the necessary conditions for the optimality of a control relative to functional (1.4) under sufficiently general constraints (1.3) imposed on the "terminal" state of the system. From it we now derive analogous conditions for certain special cases.

Let us select the functional

$$(1.16) \quad S = \int_0^X \int_0^Y f_0(x, y, z, z_x, z_y, v) dy dx$$

as the optimality criterion, where f_0 is a function of the same type as the f_i in (1.1). The terminal state conditions of the system are given by relations (1.3').

In order to apply Theorem 1 we introduce an auxiliary function by means of the equation

$$(1.17) \quad z_{0xy} = f_0(x, y, z_x, z_y, v)$$

with the supplementary conditions

$$(1.18) \quad z_0(0, y) = z_0(x, 0) = 0.$$

Then functional (1.16) can be written as $S_0 = z_0(X, Y)$. It is defined on the functions z_0, z_1, \dots, z_n given by (1.1) and (1.17) with the supplementary conditions (1.2) and (1.18). In the case being considered, H by definition has the form

$$H(x, y, U, v) = \sum_{i=0}^n u_i f_i(x, y, z, z_x, z_y, v),$$

and relations (1.6)–(1.11) yield

$$u_{0xy} = 0, \quad u_{ixy} = \frac{\partial H}{\partial z_i} - \frac{d}{dx} \left(\frac{\partial H}{\partial z_{ix}} \right) - \frac{d}{dy} \left(\frac{\partial H}{\partial z_{iy}} \right),$$

$$0 \leq x \leq X, \quad 0 \leq y \leq Y, \quad i = 1, \dots, n,$$

$$u_{0x}(x, Y) = 0, \quad u_{ix}(x, Y) + \frac{\partial H}{\partial z_{iy}} = 0 \quad \text{when } y = Y,$$

$$\begin{aligned}
u_{0y}(X, y) &= 0, \quad u_{iy}(X, y) + \frac{\partial H}{\partial z_{ix}} = 0 \quad \text{when } x = X, \\
u_0(X, Y) &= -1, \quad u_i(X, Y) + p_i = 0, \quad i = 1, \dots, n, \\
\int_0^Y &\left[f_0(X, y, z(X, y), z_x(X, y), z_y(X, y), v) \right. \\
&\quad \left. + \sum_{\alpha=1}^n p_\alpha f_\alpha(X, y, z(X, y), z_x(X, y), z_y(X, y), v) \right] dy = 0, \\
\int_0^X &\left[f_0(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v) \right. \\
&\quad \left. + \sum_{\alpha=1}^n p_\alpha f_\alpha(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v) \right] dx = 0.
\end{aligned}$$

Hence we get that $u_0(x, y) \equiv -1$, $p_\alpha = -u_\alpha(X, Y)$. Thus, the auxiliary parameters are eliminated. We obtain the following theorem.

THEOREM 2. *In order that an admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transferring the system from the state (1.2) to the state (1.3'), and the corresponding solution $z(x, y)$ of problem (1.1)–(1.2) be optimal relative to functional (1.16), it is necessary that there exist a function $u(x, y)$ such that:*

(i) *the functions $z(x, y)$, $u(x, y)$ and $v(x, y)$ form the solution of the equations*

$$\begin{aligned}
(1.19) \quad z_{ixy} &= \frac{\partial H}{\partial z_i}, \\
u_{ixy} &= \frac{\partial H}{\partial z_i} - \frac{d}{dx} \left(\frac{\partial H}{\partial z_{ix}} \right) - \frac{d}{dy} \left(\frac{\partial H}{\partial z_{iy}} \right), \quad 0 \leq x \leq X, \quad 0 \leq y \leq Y,
\end{aligned}$$

with boundary conditions

$$\begin{aligned}
(1.20) \quad z_i(0, y) &= \phi_i(y), \quad z_i(x, 0) = \psi_i(x), \quad i = 1, \dots, n, \\
u_{ix}(x, Y) + \frac{\partial H}{\partial z_{iy}} &= 0 \quad \text{when } y = Y, \\
u_{iy}(X, y) + \frac{\partial H}{\partial z_{ix}} &= 0 \quad \text{when } x = X,
\end{aligned}$$

where

$$(1.21) \quad H = \sum_{i=1}^n u_i f_i(x, y, z, z_x, z_y, v) - f_0(x, y, z, z_x, z_y, v);$$

(ii) *the conditions*

$$\begin{aligned}
 (1.22) \quad & \int_0^Y \left[\sum_{i=1}^n u_i(X, Y) f_i(X, y, z(X, y), z_x(X, y), z_y(X, y), v(X, y)) \right. \\
 & \quad \left. - f_0(X, y, z(X, y), z_x(X, y), z_y(X, y), v(X, y)) \right] dy = 0, \\
 & \int_0^X \left[\sum_{i=1}^n u_i(X, Y) f_i(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v(x, Y)) \right. \\
 & \quad \left. - f_0(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v(x, Y)) \right] dx = 0
 \end{aligned}$$

be satisfied at the point (X, Y) ;

(iii) the control $v(x, y)$ satisfies a maximum condition relative to the function $u(x, y)$.

If in the case under study the terminal state of the system is given by the conditions

$$(1.3''') \quad \int_0^X z_i(x, Y) dx = C_i, \quad i = 1, \dots, n, \quad C_i = \text{const.},$$

in which X is fixed, then the functions $u_i(x, y)$ will be determined by (1.19), but instead of (1.20) we will obtain the conditions

$$\begin{aligned}
 (1.23) \quad & z_i(0, y) = \phi_i(y), \quad z_i(x, 0) = \psi_i(x), \quad i = 1, \dots, n, \\
 & u_{ix}(x, Y) + \frac{\partial H}{\partial z_{iz}} = -q_i \quad \text{when } y = Y, \\
 & u_{iy}(X, y) + \frac{\partial H}{\partial z_{iy}} = 0 \quad \text{when } x = X, \\
 & u_i(X, Y) = 0.
 \end{aligned}$$

Condition (1.11) takes the form

$$\begin{aligned}
 (1.24) \quad & \left[\int_0^X \sum_{i=1}^n q_i \frac{\partial z_i(x, Y)}{\partial Y} \right. \\
 & \quad \left. + f_0(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v(x, Y)) \right] dx = 0.
 \end{aligned}$$

We thus obtain the following theorem.

THEOREM 3. *In order that an admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transferring the system from the state (1.2) to the state (1.3'''), and the corresponding solution $z(x, y)$ of problem (1.1)–(1.2) be optimal relative to functional (1.16) for fixed X , it is necessary that there exist functions $u_i(x, y)$ and constants q_i such that:*

(i) the functions $z(x, y)$, $u(x, y)$, $v(x, y)$ and the vector q form the solution of (1.19) with supplementary conditions (1.2) and (1.23);

(ii) the control $v(x, y)$ satisfies a maximum condition relative to the function $u(x, y)$;

(iii) condition (1.24) is satisfied at the point (X, Y) .

Example. Let the controlled process be described by the equation

$$(a) \quad z_{xt} = v - 2z_x - z_t - 2z, \quad 0 \leq x \leq X, \quad 0 \leq t \leq T,$$

where X is a fixed number. As the admissible controls we shall take the piecewise-continuous functions $v(t)$ for which the inequality $|v| \leq 1$ is satisfied. We are required to find a control which would transfer the system from the state

$$(b) \quad z(0, t) = z(x, 0) = 0$$

to the state

$$(c) \quad \int_0^X z(x, T) dx = a = \text{const.}$$

in the shortest time.

In this problem $f_0 \equiv 1$ and, consequently, $H = u(v - 2z_x - z_t - 2z) - 1$, while the equation for determining the function $u(x, t)$ takes the form

$$u_{xt} - 2u_x - u_t + 2u = 0.$$

The supplementary conditions (1.23) yield

$$u_x(x, T) - u(x, T) = -q,$$

$$u_t(X, t) - 2u(X, t) = 0,$$

$$u(X, T) = 0.$$

Hence we get

$$u(x, t) = qe^{2(t-T)}(1 - e^{x-X}).$$

From the maximum condition we find that

$$v(t) = \text{sgn} \int_0^X u(x, t) dx = \text{sgn } q.$$

We consider the two cases possible.

Case 1. $q > 0$. Then $v(t) = 1$ and, consequently, $z(x, t) = \frac{1}{2}(1 - e^{-2t}) \cdot (1 - e^{-x})$. From conditions (c) and (1.24) we get

$$(d) \quad \begin{aligned} T &= -\frac{1}{2} \ln [1 - 2a(X + e^{-X} - 1)^{-1}], \\ q &= Xe^{2T}(1 - X - e^{-X})^{-1}. \end{aligned}$$

Thus, the problem can have a solution if

$$(e) \quad 0 < 2a(X + e^{-X} - 1)^{-1} < 1.$$

Case 2. $q < 0$. Then $v(t) = -1$ and $z(x, t) = -(1 - e^{-2t})(1 - e^{-x})/2$, while from conditions (c) and (1.24) we get

$$(f) \quad \begin{aligned} T &= -\frac{1}{2} \ln [1 + 2a(X + e^{-X} - 1)^{-1}], \\ q &= Xe^{2T}(X + e^{-X} - 1)^{-1}. \end{aligned}$$

The problem can have a solution if

$$(g) \quad 0 < 1 + 2a(X + e^{-X} - 1)^{-1} < 1.$$

Thus, Theorem 3 determines the desired control in the form $v(t) = 1$ or $v(t) = -1$ depending on whether condition (e) or (g) is fulfilled. The time T corresponding to this control is determined by formula (d) or (f). However, from the arguments presented it does not follow that the controls obtained are optimal, because Theorem 3 gives only the necessary optimality conditions and the existence of an optimal control has not been proved.

Finally, let us consider the problem of minimizing functional (1.16) under the condition that the "terminal" state of the system is given in the form

$$(1.3''') \quad z_i(x, Y) = z_i^1(x), \quad z_i(X, y) = z_i^2(y), \quad i = 1, \dots, n,$$

where the $z_i^k(x)$ are given functions, piecewise-continuously differentiable and satisfying the conditions

$$(1.25) \quad z_i^1(0) = \phi_i^1(Y), \quad z_i^2(0) = \phi_i^2(X), \quad z_i^1(X) = z_i^2(Y).$$

Then, from conditions (1.7)–(1.11) we get

$$(1.26) \quad \begin{aligned} u_{ix}(x, Y) + \frac{\partial H}{\partial z_{ix}} &= -B_i(x) \quad \text{when } y = Y, \\ u_{iy}(X, y) + \frac{\partial H}{\partial z_{iy}} &= -C_i(y) \quad \text{when } x = X, \\ u_i(X, Y) &= 0, \end{aligned}$$

$$\begin{aligned} \int_0^Y \left\{ f_0(X, y, z(X, y), z_x(X, y), z_y(X, y), v(X, y)) + \sum_{i=1}^n C_i(y) z_{iy}(X, y) \right\} dy &= 0, \\ \int_0^X \left\{ f_0(x, Y, z(x, Y), z_x(x, Y), z_y(x, Y), v(x, Y)) + \sum_{i=1}^n B_i(x) z_{ix}(x, Y) \right\} dx &= 0. \end{aligned}$$

The last two equations can be simplified and take the form

$$\begin{aligned}
 & \sum_{i=1}^n \phi_{iy}^1(Y) u_i(0, Y) - \int_0^X \left[\sum_{i=1}^n \frac{\partial H(x, Y, U(x, Y), v(x, Y))}{\partial z_{iy}} z_{iy}(x, Y) \right. \\
 & \quad \left. - H(x, Y, U(x, Y), v(x, Y)) \right] dx = 0, \\
 (1.27) \quad & \sum_{i=1}^n \phi_{ix}^2(X) u_i(X, 0) - \int_0^Y \sum_{i=1}^n \frac{\partial H(X, y, U(X, y), v(X, y))}{\partial z_{ix}} z_{ix}(X, y) \\
 & \quad - H(X, y, U(X, y), v(X, y)) dy = 0.
 \end{aligned}$$

We thus obtain the following theorem.

THEOREM 4. *In order that an admissible control $v(x, y)$, $0 \leq x \leq X$, $0 \leq y \leq Y$, transferring the system from the state (1.2) to the state (1.3'''), and the corresponding solution $z(x, y)$ of problem (1.1)–(1.2) be optimal relative to functional (1.16), it is necessary that there exist functions $u_i(x, y)$, $B_i(x)$ and $C_i(y)$ such that:*

- (i) *the functions $z(x, y)$, $u(x, y)$, $B(x)$, $C(y)$, $v(x, y)$ form the solution of (1.19) with supplementary conditions (1.2) and (1.26);*
- (ii) *condition (1.27) is satisfied at the point (X, Y) ;*
- (iii) *the function $v(x, y)$ satisfies a maximum condition relative to $u(x, y)$.*

Although this theorem (as also Theorems 2 and 3) gives a complete system of relations for seeking the optimal control, its practical application calls for the solving of problems more difficult than those encountered when applying Theorems 2 and 3. Even in the simplest cases there appear complicated equations which do not always have solutions.

We can convince ourselves of this by considering the example presented above in which the terminal state of the system is given by the equation

$$(h) \quad z(x, T) = f(x),$$

where $f(x)$ is a known piecewise-continuously-differentiable function, while the admissible controls are piecewise-continuous functions $v(x, t)$ with a finite number of lines of discontinuity, satisfying the condition $|v| \leq 1$.

Then, conditions (1.26) take the form

$$\begin{aligned}
 u_x(x, T) - u(x, T) &= B(x), \\
 u_i(X, t) - 2u(X, t) &= 0, \\
 u(X, T) &= 0.
 \end{aligned}$$

Consequently,

$$\begin{aligned}
 u(x, t) &= -e^{2(t-T)} \int_x^X B(\xi) e^{x-\xi} d\xi, \\
 v(x, t) &= -\operatorname{sgn} C(x),
 \end{aligned}$$

which are defined on the function $z(x, y) = \{z_1(x, y), \dots, z_n(x, y)\}$. Let us assume that the rank of the matrix A at the point $x = X, y = Y$ equals $j + \nu + m$, and that the ranks of the matrices $B(x, Y)$ and $C(X, y)$ equal l and k , respectively. To be specific we shall consider that the determinants formed by the elements in the upper rows of these matrices are nonzero.

Then, p_α, q_δ and r_ϵ are uniquely determined by the first $j + \nu + m$ relations in (1.9). Analogously, $B_\gamma(x)$ and $C_\beta(y)$ are determined from (1.7) and (1.8). As a result, all the quantities $p_\alpha, q_\delta, r_\epsilon, B_\gamma$ and C_β are eliminated from conditions (1.7)–(1.11); moreover, the number of relations in these conditions is lessened by precisely the number of these quantities.

1.3. Estimate of the increment of functional J . In the proof of Theorem 1 which is to follow we shall require an estimate of the increment of the functional (1.13) which was used to formulate the maximum condition. In order to obtain the necessary inequalities we consider the functional

$$\begin{aligned} I[v] = & S + \int_0^X \int_0^Y \left[\sum_{i=1}^n u_i z_{ixy} - H(x, y, U, v) \right] dy dx \\ & + \sum_{\delta=1}^m q_\delta \left[\int_0^X \Phi_\delta(x, Y, z(x, Y), z_x(x, Y)) dx - a_\delta \right] \\ & + \sum_{\epsilon=1}^\nu r_\epsilon \left[\int_0^Y \Psi_\epsilon(X, y, z(X, y), z_y(X, y)) dy - b_\epsilon \right] \\ & + \int_0^X \sum_{\gamma=1}^l B_\gamma(x) \psi_\gamma^2(x, Y, z(x, Y)) dy \\ & + \int_0^Y \sum_{\beta=1}^k C_\beta(y) \psi_\beta^1(X, y, z(X, y)) dy + \sum_{\alpha=1}^j p_\alpha F_\alpha(X, Y, z(X, Y)), \end{aligned}$$

where S is the functional given by (1.4), the functions F, Φ, Ψ and ψ are taken from conditions (1.3), and H is defined by (1.12).

Let us take an arbitrary admissible control $v(x, y), 0 \leq x \leq X, 0 \leq y \leq Y$, which transfers the system from the state (1.2) to the state (1.3). We denote the solution of the corresponding Goursat problem (1.1)–(1.2) by $z(x, y, v)$. Then, $I[v(x, y)] = S$ no matter what the functions $u(x, y), B(x), C(y)$ and the constants $p_\alpha, q_\delta, r_\epsilon$ are. If with every admissible control $v(x, y)$ transferring the system from the state (1.2) to the state (1.3), we associate a solution of problem (1.1)–(1.2) for $v = v(x, y)$, then at these controls the functional I will attain its own minimal value simultaneously with S independently of the choice of the quantities u, B, C, p, q and r . In what follows we shall choose these quantities such that conditions (1.6)–(1.11) are fulfilled.

Further, let $v_1(x, y), 0 \leq x \leq X_1, 0 \leq y \leq Y_1$, be a control different

from v , which also transfers the system from the state (1.2) to the state (1.3), and let conditions (1.3) be fulfilled at the point (X_1, Y_1) . We denote the intersection of the regions D_1 , $0 \leq x \leq X_1$, $0 \leq y \leq Y_1$, and D , $0 \leq x \leq X$, $0 \leq y \leq Y$, by G , $0 \leq x \leq a$, $0 \leq y \leq b$. Both the controls v and v_1 , as well as the corresponding functions $z(x, y, v)$, $u(x, y, v)$ and $z(x, y, v_1)$, $u(x, y, v_1)$, are defined in region G . If $v(x, y)$ is an optimal control, then according to what has been said above the inequality

$$(1.28) \quad \Delta I = I[v_1] - I[v] \geq 0$$

is valid. Let us transform each term in this inequality by taking for definiteness $a = X \leq X_1$, $b = Y_1 \leq Y$.

Since the functions $z_i(x, y, v)$ and $z_i(x, y, v_1)$ have piecewise-continuous derivatives z_{ix} and z_{iy} , by the finite-increment formula,

$$(1.29) \quad \begin{aligned} \Delta S &= \sum_{i=1}^n A_i [z_i(X_1, Y_1, v_1) - z_i(X, Y, v)] \\ &= \sum_{i=1}^n A_i \{z_{ix}(X, Y, v) \Delta X + z_{iy}(X, Y, v) \Delta Y + \Delta_v z_i(a, b)\} - \eta_1, \end{aligned}$$

where

$$(1.30) \quad \begin{aligned} \Delta X &= X_1 - X, \quad \Delta Y = Y_1 - Y, \\ \Delta_v z_i(a, b) &= z_i(a, b, v_1) - z_i(a, b, v), \\ \eta_1 &= \sum_{i=1}^n A_i \left\{ \sum_{k=1}^q [z_{ix}(X, Y, v) - z_{iy}(\bar{x}_k, b, v_1)] \Delta x_k \right. \\ &\quad \left. + \sum_{j=1}^r [z_{iy}(X, Y, v) - z_{iy}(a, \bar{y}_j, v)] \Delta y_j \right\}. \end{aligned}$$

Here x_1, \dots, x_q and y_1, \dots, y_r are the collections of all the points of discontinuity of the functions $z_{ix}(x, Y, v)$ and $z_{iy}(X, y, v_1)$ in the region $X \leq x \leq X_1$, $Y_1 \leq y \leq Y$, and $\Delta x_k = x_k - x_{k-1}$, $\Delta y_j = y_j - y_{j-1}$.

Analogously we find

$$(1.31) \quad \begin{aligned} &\sum_{\alpha=1}^j p_\alpha [F_\alpha(X_1, Y_1, z(X_1, Y_1, v_1)) - F_\alpha(X, Y, z(X, Y, v))] \\ &= \sum_{\alpha=1}^j p_\alpha \left[\frac{dF_\alpha(X, Y, z(X, Y, v))}{dX} \Delta X + \frac{dF_\alpha(X, Y, z(X, Y, v))}{dY} \Delta Y \right. \\ &\quad \left. + \sum_{i=1}^n \left[\frac{\partial F_\alpha(X, Y, z(X, Y, v))}{\partial z_i} \Delta_v z_i(a, b) \right] \right] - \eta_2, \end{aligned}$$

where

$$\begin{aligned}
 \eta_2 = \sum_{\alpha=1}^j p_{\alpha} \left\{ \sum_{k=1}^q \left[\frac{dF_{\alpha}(X, Y, z(X, Y, v))}{dX} - \frac{dF_{\alpha}(\bar{x}_k, b, z(\bar{x}, b, v_1))}{dX} \right] \Delta x_k \right. \\
 (1.32) \quad \left. + \sum_{s=1}^r \left[\frac{dF_{\alpha}(X, Y, z(X, Y, v))}{dY} - \frac{dF_{\alpha}(X, \bar{y}_s, z(X, \bar{y}_s, v))}{dY} \right] \Delta y_s \right. \\
 \left. + \sum_{i=1}^n \left[\frac{\partial F_{\alpha}(X, Y, z(X, Y, v))}{\partial z_i} - \frac{\partial F_{\alpha}(a, b, \bar{z})}{\partial z_i} \Delta_v z_i(a, b) \right] \right\}.
 \end{aligned}$$

It should be noted that different magnitudes of p_{α} may correspond to the different controls v and v_1 . However, when constructing the increment of functional I we need not consider terms containing the increment Δp_{α} since the coefficients of $p_{\alpha}(v_1)$ and of $p_{\alpha}(v)$ are zero according to conditions (1.3₁). By the very same reasoning we need not pay any attention to the terms containing $\Delta \psi_i^k$, Δq_{δ} and Δr_{ϵ} .

Let I_0 denote the double integral in functional I . Then, taking into account that the z_i form the solution of (1.1) and that $G = D_1 \times D$, we have

$$\begin{aligned}
 \Delta I_0 &= \iint_{D_1} \left[\sum_{i=1}^n u_i(x, y, v_1) z_{ixy}(x, y, v_1) - H(x, y, U(x, y, v_1), v_1) \right] dy \, dx \\
 (1.33) \quad &- \iint_D \left[\sum_{i=1}^n u_i(x, y, v) z_{ixy}(x, y, v) - H(x, y, U(x, y, v), v) \right] dy \, dx \\
 &= \iint_G \left[\sum_{i=1}^n (\Delta u_i \Delta z_{ixy} + u_i \Delta z_{ixy} + \Delta u_i z_{ixy}) - \Delta H \right] dy \, dx,
 \end{aligned}$$

where

$$\begin{aligned}
 \Delta H &= H(x, y, U(x, y, v_1), v_1) - H(x, y, U(x, y, v), v), \\
 \Delta U_i &= U_i(x, y, v_1) - U_i(x, y, v).
 \end{aligned}$$

From the definitions of the functions Δz_i and Δu_i , it follows that they satisfy the equations

$$(1.34) \quad \Delta z_{ixy} = \Delta \frac{\partial H}{\partial u_i}, \quad \Delta u_{ixy} = \Delta \frac{\partial H}{\partial z_i} - \frac{d}{dx} \left(\Delta \frac{\partial H}{\partial z_{ix}} \right) - \frac{d}{dy} \left(\Delta \frac{\partial H}{\partial z_{iy}} \right),$$

where

$$\Delta \frac{\partial H}{\partial U_i} = \frac{\partial H(x, y, U(x, y, v_1), v_1)}{\partial U_i} - \frac{\partial H(x, y, U(x, y, v), v)}{\partial U_i}.$$

Furthermore, the equalities

$$\Delta z_i(0, y) = \Delta z_i(x, 0) = 0$$

are valid according to conditions (1.2).

Since Green's formula

$$\iint_G [qp_{xy} - pq_{xy}] dy dx = [(pq)_{x=0}]_{y=0}^b - \int_0^a [pq_x]_{y=0}^b dx - \int_0^b [pq_y]_{x=0}^a dy$$

is valid for any continuous functions p and q having piecewise-continuous derivatives p_x , p_y , p_{xy} , q_x , q_y and q_{xy} , we shall have

$$(1.35) \quad \begin{aligned} & \iint_G \sum_{i=1}^n \Delta u_i \Delta z_{ixy} dy dx \\ &= \iint_G \sum_{i=1}^n \left(\Delta \frac{\partial H}{\partial z_i} \Delta z_i + \Delta \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \Delta \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right) dy dx - 2\eta_3, \end{aligned}$$

where

$$(1.36) \quad \begin{aligned} \eta_3 = & -\frac{1}{2} \sum_{i=1}^n \left\{ \Delta_v u_i(a, b) \Delta_v z_i(a, b) \right. \\ & \left. - \int_0^a \left[\left(\Delta u_{ix} + \Delta \frac{\partial H}{\partial z_{ix}} \right) \Delta z_i \right]_{y=b} dy - \int_0^b \left[\left(\Delta u_{iy} + \Delta \frac{\partial H}{\partial z_{iy}} \right) \Delta z_i \right]_{x=a} dy \right\}. \end{aligned}$$

On the other hand, with due regard to (1.34) we have

$$(1.37) \quad \iint_G \sum_{i=1}^n \Delta u_i \Delta z_{ixy} dy dx = \iint_G \sum_{i=1}^n \Delta \frac{\partial H(x, y, U(x, y, v), v)}{\partial u_i} \Delta u_i dy dx.$$

From (1.35) and (1.37) it follows that

$$(1.38) \quad \begin{aligned} & \iint_G \sum_{i=1}^n \Delta u_i \Delta z_{ixy} dy dx \\ &= \frac{1}{2} \iint_G \sum_{i=1}^{4n} \Delta \frac{\partial H(x, y, U(x, y, v), v)}{\partial U_i} \Delta U_i dy dx - \frac{1}{2} \eta_3. \end{aligned}$$

Analogously we find

$$(1.39) \quad \begin{aligned} & \iint_G \sum_{i=1}^n u_i \Delta z_{ixy} dy dx = \sum_{i=1}^n \left\{ u_i(X, Y, v) \Delta_v z_i(a, b) \right. \\ & \quad - \int_0^a \left(u_{ix} + \frac{\partial H}{\partial z_{ix}} \right)_{y=Y} \Delta_v z_i(x, b) dx \\ & \quad - \int_0^b \left(u_{iy} + \frac{\partial H}{\partial z_{iy}} \right)_{x=X} \Delta_v z_i(a, y) dy \\ & \quad \left. + \iint_G \left(\frac{\partial H}{\partial z_i} \Delta z_i + \frac{\partial H}{\partial z_{ix}} \Delta z_{ix} + \frac{\partial H}{\partial z_{iy}} \Delta z_{iy} \right) dy dx \right\} - \eta_4, \\ & \iint_G \sum_{i=1}^n \Delta u_i z_{ixy} dy dx = \iint_G \sum_{i=1}^n \frac{\partial H}{\partial u_i} \Delta u_i dy dx, \end{aligned}$$

where

$$\begin{aligned}
 \eta_4 = \sum_{i=1}^n & \left\{ [u_i(X, Y, v) - u_i(a, b, v)] \Delta_v z_i(a, b) \right. \\
 (1.40) \quad & - \int_0^a \left(u_{ix} + \frac{\partial H}{\partial z_{iy}} \right)_{y=b}^x \Delta_v z_i(x, b) dx \\
 & \left. - \int_0^b \left(u_{iy} + \frac{\partial H}{\partial z_{ix}} \right)_{x=a}^y \Delta_v z_i(a, y) dy \right\}.
 \end{aligned}$$

Taking (1.38) and (1.39) into account we get

$$\begin{aligned}
 & \iint_G \left\{ \sum_{i=1}^n (\Delta u_i \Delta z_{ixy} + u_i \Delta z_{ixy} + \Delta u_i z_{ixy}) - \Delta H \right\} dy dx \\
 & = \sum_{i=1}^n \left\{ u_i(X, Y) \Delta_v z_i(a, b) - \int_0^a \left(u_{ix} + \frac{\partial H}{\partial z_{iy}} \right)_{y=Y}^x \Delta_v z_i(x, b) dx \right. \\
 & \quad - \int_0^b \left(u_{iy} + \frac{\partial H}{\partial z_{ix}} \right)_{x=X}^y \Delta_v z_i(a, y) dy \\
 & \quad \left. - \iint_G \left[\Delta H - \sum_{i=1}^{4n} \left(\frac{\partial H}{\partial U_i} + \frac{1}{2} \Delta \frac{\partial H}{\partial U_i} \right) \Delta U_i \right] dy dx \right\} - \eta_3 - \eta_4.
 \end{aligned}$$

Applying Taylor's formula to the functions H and $\partial H / \partial w_i$ and restricting ourselves in the expansions to second order terms in Δw_i , we have

$$\begin{aligned}
 & \iint_G \left\{ \sum_{i=1}^n (\Delta u_i \Delta z_{ixy} + u_i \Delta z_i + \Delta u_i z_{ixy}) - \Delta H \right\} dy dx \\
 (1.41) \quad & = \sum_{i=1}^n \left\{ u_i(X, Y) \Delta_v z_i(a, b) - \int_0^a \left(u_{ix} + \frac{\partial H}{\partial z_{iy}} \right)_{y=Y}^x \Delta_v z_i(x, b) dx \right. \\
 & \quad - \int_0^b \left(u_{iy} + \frac{\partial H}{\partial z_{ix}} \right)_{x=X}^y \Delta_v z_i(a, y) dy \Big\} \\
 & \quad - \iint_G \Delta_v H dy dx - \eta_3 - \eta_4 - \eta_5,
 \end{aligned}$$

where

$$\begin{aligned}
 \Delta_v H & = H(x, y, U(x, y, v), v_1) - H(x, y, U(x, y, v), v), \\
 \eta_5 & = \frac{1}{2} \iint_G \sum_{i=1}^{4n} \left[\frac{\partial H(x, y, U(x, y, v), v_1)}{\partial U_i} - \frac{\partial H(x, y, U(x, y, v), v)}{\partial U_i} \right] \\
 (1.42) \quad & + \sum_{j=1}^{4n} \left[\frac{\partial^2 H(x, y, U(x, y, v) + \theta_1 \Delta U, v_1)}{\partial U_i \partial U_j} \right. \\
 & \quad \left. - \frac{\partial^2 H(x, y, U(x, y, v) + \theta_2 \Delta U, v)}{\partial U_i \partial U_j} \Delta U_j \right] \Delta U_i dy dx, \\
 & \quad 0 \leq \theta_i \leq 1.
 \end{aligned}$$

With due regard to the notation of (1.13), from (1.33) and (1.41) we get

$$\begin{aligned}
 \Delta I_0 = & -\Delta J + \sum_{i=1}^n \left\{ u_i(X, Y) \Delta_v z_i(a, b) \right. \\
 (1.43) \quad & - \int_0^a \left(u_{ix} + \frac{\partial H}{\partial z_{iy}} \right)_{y=Y} \Delta_v z_i(x, b) dx \\
 & \left. - \int_0^b \left(u_{iy} + \frac{\partial H}{\partial z_{ix}} \right)_{x=X} \Delta_v z_i(a, y) dy \right\} - \eta_3 - \eta_4 - \eta_5.
 \end{aligned}$$

Let us now transform those terms in inequality (1.28) which correspond to conditions (1.3₃).

The functions $z_i(x, y, v_1)$ and $z_i(x, y, v)$ satisfy conditions (1.3₃). Therefore,

$$\begin{aligned}
 (1.44) \quad & \sum_{\delta=1}^m q_\delta \left\{ \left[\int_0^{X_1} \Phi_\delta(x, Y_1, z(x, Y_1, v_1), z_x(x, Y_1, v_1)) dx - a_\delta \right] \right. \\
 & \left. - \left[\int_0^X \Phi_\delta(x, Y, z(x, Y, v), z_x(x, Y, v)) dx - a_\delta \right] \right\} \\
 & = \sum_{\delta=1}^m q_\delta \left\{ \sum_{i=1}^n \frac{\partial \Phi_\delta(X, Y, w(X, Y, v))}{\partial z_{ix}} \Delta_v z_i(a, b) \right. \\
 & + \Phi_\delta(X, Y, w(X, Y, v)) \Delta X + \int_0^X \left[\frac{d\Phi_\delta(x, Y, w(x, Y, v))}{dY} \Delta Y \right. \\
 & + \sum_{i=1}^n \left(\frac{\partial \Phi_\delta(x, Y, w(x, Y, v))}{\partial z_i} \right. \\
 & \left. \left. - \frac{d}{dx} \left[\frac{\partial \Phi_\delta(x, Y, w(x, Y, v))}{\partial z_{ix}} \right] \right) \Delta_v z_i(x, b) \right] dx \Big\} - \eta_6 - \eta_7,
 \end{aligned}$$

where, as above, it is assumed that $X = a \leq X_1$, $Y_1 = b \leq Y$, w denotes the vector with components z_i and z_{ix} , and

$$\begin{aligned}
 (1.45) \quad \eta_6 = & \sum_{\delta=1}^m q_\delta \left\{ [\Phi_\delta(X, Y, w(X, Y, v) - \Phi_\delta(\bar{x}, Y, w(\bar{x}, Y, v_1))] \Delta X \right. \\
 & + \int_0^X \left[\frac{d\Phi_\delta(x, Y, w(x, Y, v))}{dY} - \frac{d\Phi_\delta(x, \bar{y}, w(x, \bar{y}, v))}{dY} \right] \Delta Y dx \Big\}, \\
 \eta_7 = & \sum_{\delta=1}^m \sum_{i=1}^{2n} q_\delta \int_0^a \left[\frac{\partial \Phi_\delta(x, b, w(x, b, v))}{\partial w_i} - \frac{\partial \Phi_\delta(x, b, \bar{w})}{\partial w_i} \right] \\
 & \cdot \Delta w_i(x, b) dx
 \end{aligned}$$

Analogously we find

$$\begin{aligned}
(1.46) \quad & \sum_{\epsilon=1}^p r_{\epsilon} \left\{ \left[\int_0^{Y_1} \Psi_{\epsilon}(X_1, y, z(X_1, y, v_1), z_y(X_1, y, v_1)) dy - b_{\epsilon} \right] \right. \\
& \left. - \left[\int_0^Y \Psi_{\epsilon}(X, y, z(X, y, v), z_y(X, y, v)) dy - b_{\epsilon} \right] \right\} \\
& = \sum_{\epsilon=1}^p r_{\epsilon} \left\{ \sum_{i=1}^n \frac{\partial \Psi_{\epsilon}(X, Y, w(X, Y, v))}{\partial z_{iy}} \Delta_v z_i(a, b) \right. \\
& + \Psi_{\epsilon}(X, Y, w(X, Y, v)) \Delta Y \\
& + \int_0^Y \left[\frac{d\Psi_{\epsilon}(X, y, w(X, y, v))}{dX} \Delta X + \sum_{i=1}^n \left(\frac{\partial \Psi_{\epsilon}(X, y, w(X, y, v))}{\partial z_i} \right. \right. \\
& \left. \left. - \frac{d}{dy} \left[\frac{\partial \Psi_{\epsilon}(X, y, w(X, y, v))}{\partial z_{iy}} \right] \right) \Delta_v z_i(a, y) \right] dy \Big\} - \eta_8 - \eta_9,
\end{aligned}$$

where w is the vector with components z_i and z_{iy} , and

$$\begin{aligned}
(1.47) \quad \eta_8 &= \sum_{\epsilon=1}^p r_{\epsilon} \left\{ \sum_{s=1}^r [\Psi_{\epsilon}(X, Y, w(X, Y, v)) - \Psi_{\epsilon}(X, \bar{y}_s, w(X, \bar{y}_s, v))] \Delta y_s \right. \\
& \quad \left. + \sum_{k=1}^q \left[\int_0^Y \frac{d\Psi_{\epsilon}(X, y, w(X, y, v))}{dX} dy \right. \right. \\
& \quad \left. \left. - \int_0^b \frac{d\Psi_{\epsilon}(\bar{x}_k, y, w(\bar{x}_k, y, v_1))}{dX} dy \right] \Delta x_k \right\}, \\
\eta_9 &= \sum_{\epsilon=1}^p r_{\epsilon} \sum_{i=1}^{2n} \int_0^b \left[\frac{\partial \Psi_{\epsilon}(a, y, w(a, y, v))}{\partial w_i} - \frac{\partial \Psi_{\epsilon}(a, y, \bar{w})}{\partial w_i} \right] \\
& \quad \cdot \Delta w_i(a, y) dy.
\end{aligned}$$

Finally, let us transform those terms in inequality (1.28) which correspond to conditions (1.3₂).

We have

$$\begin{aligned}
(1.48) \quad & \sum_{\beta=1}^k \left[\int_0^{Y_1} C_{\beta}(y) \psi_{\beta}^1(X_1, y, z(X_1, y, v_1)) dy \right. \\
& \left. - \int_0^Y C_{\beta}(y) \psi_{\beta}^1(X, y, z(X, y, v)) dy \right] \\
& = \sum_{\beta=1}^k \left\{ \int_0^Y C_{\beta}(y) \frac{d\psi_{\beta}^1(X, y, z(X, y, v))}{dX} \Delta X dy \right. \\
& \quad \left. + \int_0^b C_{\beta}(y) \sum_{i=1}^n \frac{\partial \psi_{\beta}^1(X, y, z(X, y, v))}{\partial z_i} \Delta_v z_i(X, y, v) dy \right\} \\
& \quad - \eta_{10} - \eta_{11},
\end{aligned}$$

where

$$\begin{aligned}
 \eta_{10} &= \sum_{\beta=1}^k \sum_{k=1}^q \left[\int_0^Y C_{\beta}(y) \frac{d\psi_{\beta}^1(X, y, z(X, y, v))}{dX} dy \right. \\
 (1.49) \quad &\quad \left. - \int_0^Y C_{\beta}(y) \frac{d\psi_{\beta}^1(\bar{x}_k, y, z(\bar{x}_k, y, v_1))}{dX} dy \right] \Delta x_k, \\
 \eta_{11} &= \sum_{\beta=1}^k \int_0^b C_{\beta}(y) \sum_{i=1}^n \left[\frac{\partial \psi_{\beta}^1(X, y, z(X, y, v))}{\partial z_i} - \frac{\partial \psi_{\beta}^1(X, y, \bar{z})}{\partial z_i} \right] \\
 &\quad \cdot \Delta_v z_i(a, y) dy.
 \end{aligned}$$

Analogously we get

$$\begin{aligned}
 &\sum_{\gamma=1}^l \left[\int_0^{X_1} B_{\gamma}(x) \psi_{\gamma}^2(x, Y_1, z(x, Y_1, v_1)) dy \right. \\
 &\quad \left. - \int_0^X B_{\gamma}(x) \psi_{\gamma}^2(x, Y, z(x, Y, v)) dy \right] \\
 (1.50) \quad &= \sum_{\gamma=1}^l \left\{ \int_0^X B_{\gamma}(x) \frac{d\psi_{\gamma}^2(x, Y, z(x, Y, v))}{dY} \Delta Y dx \right. \\
 &\quad \left. + \int_0^a B_{\gamma}(x) \sum_{i=1}^n \frac{\partial \psi_{\gamma}^2(x, Y, z(x, Y, v))}{\partial z_i} \Delta_v z_i(x, b, v) dx \right\} \\
 &\quad - \eta_{12} - \eta_{13},
 \end{aligned}$$

where

$$\begin{aligned}
 \eta_{12} &= \sum_{s=1}^r \sum_{\gamma=1}^l \int_0^X B_{\gamma}(x) \left[\frac{d\psi_{\gamma}^2(x, Y, z(x, Y, v))}{dY} \right. \\
 &\quad \left. - \frac{d\psi_{\gamma}^2(x, \bar{y}_s, z(x, \bar{y}_s, v))}{dY} \right] \Delta y_s dx, \\
 (1.51) \quad \eta_{13} &= \sum_{\gamma=1}^l \int_0^a B_{\gamma}(x) \sum_{i=1}^n \left[\frac{d\psi_{\gamma}^2(x, Y, z(x, Y, v))}{dY} \right. \\
 &\quad \left. - \frac{\partial \psi_{\gamma}^2(x, Y, \bar{z})}{\partial z_i} \right] \Delta_v z_i(x, b) dx.
 \end{aligned}$$

If now we reckon that the functions $u(x, y)$, $B(x)$ and $C(y)$ together with the constants p_{α} , q_{δ} and r_{ϵ} transform the system of equations (1.6)–(1.11) into an identity, then from relations (1.28), (1.29), (1.43), (1.44), (1.46), (1.48) and (1.50) we obtain

$$\begin{aligned}
 (1.52) \quad -\Delta J[v] &= - \iint_G [H(x, y, U(x, y, v), v_1) \\
 &\quad - H(x, y, U(x, y, v), v)] dy dx \geq \eta,
 \end{aligned}$$

where $n = \sum_{i=1}^{13} \eta_i$. This inequality permits us to obtain an estimate of the increment of functional J in terms of the increments of the quantities X , Y , and v . Here, the right-hand side of inequality (1.52) should be evaluated beforehand.

As has already been noted above, the functions Δz_i and Δu_i form the solution of (1.34) and, moreover, the $\Delta z_i(x, y)$ satisfy the trivial boundary conditions: $\Delta z_i(0, y) = \Delta z_i(x, 0) = 0$. In [9] it has been shown that for functions thus defined the inequalities

$$\begin{aligned}
 |\Delta z_i(x, y)| &\leq Q \iint_G \Delta v(x, y) dy dx, \\
 |\Delta z_{ix}(x, y)| &\leq Q_1 \iint_G \Delta v(x, y) dy dx + R_1 \int_0^b \Delta v(a, y) dy, \\
 |\Delta z_{iy}(x, y)| &\leq Q_2 \iint_G \Delta v(x, y) dy dx + R_2 \int_0^a \Delta v(x, b) dx, \\
 0 &\leq x \leq a, \quad 0 \leq y \leq b, \quad i = 1, \dots, n',
 \end{aligned}
 \tag{1.53}$$

where Q_i and R_i are positive constants and $\Delta v = \sum_{r=1}^r |\Delta v_r(x, y)|$ (r is the dimension of the control region), are valid.

The functions $u_i(x, y, v)$, defined in region G , satisfy the linear equations (1.6). The functions $u_i(x, Y, v)$ and $u_i(X, y, v)$ satisfy the ordinary linear differential equations (1.7) and (1.8) on the boundary of this region. Their initial values at the point (X, Y) are determined from (1.9):

$$(1.54) \quad u_i(x, Y, v)|_{x=X} = \varepsilon_i(v), \quad u_i(X, y, v)|_{y=Y} = \varepsilon_i(v),$$

where the $\varepsilon_i(v)$ denote the right-hand sides of relations (1.9). If we let $\{u_{ij}(x, X, Y, v)\}$ denote the transition matrix (normal when $x = X$) of the system (1.7), then the solution of the Cauchy problem (1.7), (1.54) can be represented in the form

$$\begin{aligned}
 u_i(x, Y, v) &= \sum_{j=1}^n \left\{ u_{ij}(x, X, Y, v) \varepsilon_j(v) \right. \\
 &\quad + \sum_{\delta=1}^m q_\delta \left[K_{ij}(x, \xi, Y, v) \frac{\partial \Phi_\delta(\xi, Y, w(\xi, Y, v))}{\partial z_{i\delta}} \right]_{\xi=x}^x \\
 &\quad - \int_x^X \left(K_{ij}(x, \xi, Y, v) \frac{\partial \Phi_\delta(\xi, Y, w(\xi, Y, v))}{\partial z_i} \right. \\
 &\quad \left. + \frac{\partial K_{ij}(x, \xi, Y, v)}{\partial \xi} \frac{\partial \Phi_\delta(\xi, Y, w(\xi, Y, v))}{\partial z_{i\delta}} \right) d\xi \Bigg] \\
 &\quad \left. + \sum \int_x^X K_{ij} \frac{\partial \psi_\gamma^2}{\partial z_j} B_\gamma(\xi) d\xi \right\},
 \end{aligned}
 \tag{1.55}$$

where w is the vector with components z_i and z_{ix} , and K_{ij} is the Cauchy function corresponding to the transition matrix $\{u_{ij}(x, X, Y, v)\}$. Analogously we find the solution of the problem (1.8), (1.54):

$$\begin{aligned}
 (1.56) \quad u_i(X, y, v) = & \sum_{i=1}^n \left\{ u_{ij}(X, y, Y, v) \varepsilon_j(v) \right. \\
 & + \sum_{\epsilon=1}^r r_{\epsilon} \left[K_{ij}(X, y, \eta, v) \frac{\partial \Psi_{\epsilon}(X, \eta, w(X, \eta, v))}{\partial z_{iy}} \right]_{\eta=y}^Y \\
 & - \int_y^Y \left(K_{ij}(X, y, \eta, v) \frac{\partial \Psi_{\epsilon}(X, \eta, w(X, \eta, v))}{\partial z_i} \right. \\
 & \left. + \frac{\partial K_{ij}(X, y, \eta, v)}{\partial \eta} \frac{\partial \Psi_{\epsilon}(X, \eta, w(X, \eta, v))}{\partial z_{iy}} \right) d\eta \left. \right] \\
 & + \sum \int_y^Y K_{ij} \frac{\partial \psi_{\beta}^1}{\partial z_j} C_{\beta}(\eta) d\eta \left. \right\},
 \end{aligned}$$

where w is the vector with components z_i and z_{iy} , and K_{ij} is the Cauchy function corresponding to the transition matrix $\{u_{ij}(X, y, Y, v)\}$ of system (1.8).

Consequently, (1.6) with the supplementary conditions (1.7)–(1.9) can now be reduced to the integral equations

$$\begin{aligned}
 (1.57) \quad u_i(x, y, v) = & u_i(X, y, v) + u_i(x, Y, v) - u_i(X, Y, v) \\
 & + \int_x^X \frac{\partial H}{\partial z_{iy}} \Big|_{y=y}^Y dx + \int_y^Y \frac{\partial H}{\partial z_{ix}} \Big|_{x=x}^X dy + \int_x^X \int_y^Y \frac{\partial H}{\partial z_i} dy dx,
 \end{aligned}$$

where $u_i(x, Y, v)$ and $u_i(X, y, v)$ are determined by relations (1.55) and (1.56). Analogously we obtain integral equations for the determination of the functions u_i corresponding to the control v_1 and to the former parameters p_{α} , q_{δ} , r_{ϵ} , B_{γ} and C_{β} :

$$\begin{aligned}
 (1.58) \quad u_i(x, y, v_1) = & u_i(X_1, y, v_1) + u_i(x, Y_1, v_1) - u_i(X_1, Y_1, v_1) \\
 & + \int_x^X \frac{\partial H}{\partial z_{iy}} \Big|_{y=y}^{Y_1} dx + \int_y^{Y_1} \frac{\partial H}{\partial z_{ix}} \Big|_{x=x}^{X_1} dy + \int_x^X \int_y^{Y_1} \frac{\partial H}{\partial z_i} dy dx.
 \end{aligned}$$

Since the functions F_{α} , Φ_{δ} and Ψ_{ϵ} are continuously differentiable in x and y and also twice continuously differentiable in all the remaining arguments, by applying the Lipschitz condition and inequalities (1.53) we obtain

$$\begin{aligned}
 & \left| \frac{\partial F_{\alpha}(X_1, Y_1, z(X_1, Y_1, v_1))}{\partial z_i} - \frac{\partial F_{\alpha}(X, Y, z(X, Y, v))}{\partial z_i} \right| \\
 & \leq M_1 \Delta X + N_1 \Delta Y + P_1 \iint_G \Delta v(x, y) dy dx,
 \end{aligned}$$

$$\begin{aligned}
 (1.59) \quad & \left| \frac{\partial \Phi_\delta(X_1, Y_1, w(X_1, Y_1, v_1))}{\partial w_i} - \frac{\partial \Phi_\delta(X, Y, w(X, Y, v))}{\partial w_i} \right| \\
 & \leq M_2 \Delta X + N_2 \Delta Y + P_2 \iint_G \Delta v(x, y) dy dx + S_1 \int_0^b \Delta v(x, y) dy, \\
 & \left| \frac{\partial \Psi_\epsilon(X_1, Y_1, w(X_1, Y_1, v_1))}{\partial w_i} - \frac{\partial \Psi_\epsilon(X, Y, w(X, Y, v))}{\partial w_i} \right| \\
 & \leq M_3 \Delta X + N_3 \Delta Y + P_3 \iint_G \Delta v(x, y) dy dx + S_2 \int_0^a \Delta v(x, b) dx,
 \end{aligned}$$

where M_i , N_i , P_i and S_i are positive constants, $\Delta X = |X_1 - X|$, $\Delta Y = |Y_1 - Y|$. Taking these inequalities into account we find that the inequalities

$$\begin{aligned}
 |\mathcal{E}_i(v_1) - \mathcal{E}_i(v)| & \leq M_4 \Delta X + N_4 \Delta Y + P_4 \iint_G \Delta v(x, y) dy dx \\
 & + S_3 \int_0^b \Delta v(a, y) dy + S_4 \int_0^a \Delta v(x, b) dx
 \end{aligned}$$

are valid for the quantities $\mathcal{E}_i(v)$ occurring in conditions (1.54).

Since the Cauchy functions are bounded in the regions D , $0 \leq x \leq X$, $0 \leq y \leq Y$, and D_1 , $0 \leq x \leq X_1$, $0 \leq y \leq Y_1$, from (1.55) and (1.56) we have

$$\begin{aligned}
 |u_i(x, Y_1, v_1) - u_i(x, Y, v)| & \leq M_5 \Delta X + N_5 \Delta Y + P_5 \iint_G \Delta v(x, y) dy dx \\
 & + S_5 \int_0^b \Delta v(a, y) dy + S_6 \int_0^a \Delta v(x, b) dx, \\
 |u_i(X_1, y, v_1) - u_i(X, y, v)| & \leq M_6 \Delta X + N_6 \Delta Y + P_6 \iint_G \Delta v(x, y) dy dx \\
 & + S_7 \int_0^b \Delta v(a, y) dy + S_8 \int_0^a \Delta v(x, b) dx
 \end{aligned}$$

for all points (x, y) of the region G , $0 \leq x \leq a$, $0 \leq y \leq b$. Analogous inequalities are satisfied by the functions $\Delta_v u_{ix}(x, b)$ and $\Delta_v u_{iy}(a, y)$. If now we apply the method used in [9] to obtain inequality (1.54), we have

$$\begin{aligned}
 (1.60) \quad & |\Delta_v u_i(x, y)| \leq M_7 \Delta X + N_7 \Delta Y + P_7 \iint_G \Delta v(x, y) dy dx \\
 & + S_9 \int_0^b \Delta v(a, y) dy + S_{10} \int_0^a \Delta v(x, b) dx.
 \end{aligned}$$

The estimates of the increments of functional J as well as the solutions of (1.1) and (1.6) which we have obtained will permit us to complete the proof of Theorem 1.

1.4. Proof of Theorem 1. Thus, let $v(x, y)$ be an optimal control which transfers the system from the state (1.2) to the state (1.3), and let $z(x, y, v)$ and $u(x, y, v)$ be the corresponding functions determined by relations (1.1)–(1.2) and (1.6)–(1.11). Then, inequality (1.52) is valid for any other admissible control $v_1(x, y)$ which also transfers the system from the state (1.2) to the state (1.3).

Let us suppose that Theorem 1 is false. Then we can find an admissible control $v_1(x, y)$ such that the inequalities

$$\begin{aligned} \Delta J[v] &= \iint_G [H(x, y, U(x, y, v), v_1) - H(x, y, U(x, y, v), v)] dy dx > 0, \\ (1.61) \quad -\Delta J[v] &= -\iint_G [H(x, y, U(x, y, v), v_1) \\ &\quad - H(x, y, U(x, y, v), v)] dy dx \geq \eta \end{aligned}$$

are satisfied simultaneously (see (1.52)). From inequalities (1.61) it follows that there exists ϵ_1 such that for some region G_ϵ , $G_\epsilon \subset G$, whose area is ϵ , $\epsilon < \epsilon_1$, the inequality

$$(1.62) \quad H(x, y, U(x, y, v), v_1) - H(x, y, U(x, y, v), v) \geq \delta > 0, \quad (x, y) \in G_\epsilon,$$

is satisfied, where δ is some positive number depending on ϵ_1 .

Let us take an auxiliary control $v_\epsilon(x, y)$ defined in the region D_ϵ , $0 \leq x \leq X_\epsilon$, $0 \leq y \leq Y_\epsilon$, and satisfying the following conditions:

- (i) it transfers the system from the state (1.2) to the state (1.3) and, moreover, condition (1.3) is satisfied at the point (X_ϵ, Y_ϵ) ;
- (ii) it is defined in the region $D \times D_\epsilon$ by the formula

$$v_\epsilon = \begin{cases} v_1 & \text{if } (x, y) \in G_\epsilon, \\ v & \text{if } (x, y) \in D \times D_\epsilon \setminus G_\epsilon, \end{cases}$$

where G_ϵ is the region in which inequality (1.62) is satisfied;

- (iii) the function $v_\epsilon(x, y)$ is assumed to be arbitrary outside the region $D \times D_\epsilon$, except that it should be admissible and should satisfy condition (i).

Because of the condition that the class of admissible controls is complete, the function $v_\epsilon(x, y)$ must be chosen so as to satisfy the inequality

$$(1.63) \quad |X - X_\epsilon| < L\epsilon, \quad |Y - Y_\epsilon| < L\epsilon,$$

where the constant L is independent of ϵ .

Since the area of region G_ϵ equals ϵ , from (1.53) we get

$$\begin{aligned} |\Delta z_i(x, y)| &\leq Q\epsilon \iint_{G_\epsilon} \Delta v^2(x, y) dy, \\ |\Delta z_{ix}(x, y)| &\leq Q_1\epsilon \iint_G \Delta v^2(x, y) dy dx, \\ |\Delta z_{iy}| &\leq Q_2\epsilon \iint_G \Delta v^2(x, y) dy dx, \end{aligned}$$

where $\Delta v = v_\epsilon - v$. It follows from the definition of admissible controls that there exists a constant B such that $|\Delta v(x, y)| \leq B$ for $(x, y) \in D \times D_\epsilon$ and, consequently, for all points (x, y) of region $D \times D_\epsilon$ we have

$$(1.64) \quad |\Delta z_i(x, y)| \leq T\epsilon^2, \quad |\Delta z_{ix}(x, y)| \leq T\epsilon^2, \quad |\Delta z_{iy}(x, y)| \leq T\epsilon^2, \\ T = \text{const.} > 0.$$

Analogously, from inequalities (1.60) and (1.63) we obtain

$$(1.65) \quad |\Delta_v u_i(x, y)| \leq T_1\epsilon, \quad |\Delta_v u_{ix}| \leq T_1\epsilon, \quad |\Delta_v u_{iy}| \leq T_1\epsilon.$$

Using the inequalities obtained, from relation (1.30) we have

$$(1.66) \quad |\eta_1| \leq T_2\epsilon^2.$$

Applying inequalities (1.59), (1.63) and (1.65) we obtain an estimate for the quantity η_2 given in (1.32),

$$(1.67) \quad |\eta_2| \leq T_3\epsilon^2.$$

We now estimate the quantity η_3 defined by (1.36). From inequalities (1.64) and (1.65) it follows that

$$|\Delta_v u_i(a, b) \Delta_v z_i(a, b)| \leq o(\epsilon^2) \quad \text{where} \quad \frac{o(\epsilon^2)}{\epsilon^2} \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0.$$

Furthermore, since the right-hand sides of (1.7) and (1.9) satisfy Lipschitz conditions in the variables z , z_x and z_y , and since the functions $B_\gamma(x)$, $C_\beta(y)$ and the constants q_δ and r_ϵ are chosen to be one and the same for the functions $u(x, y, v)$ and $u(x, y, v_1)$, according to inequalities (1.64) we have

$$\begin{aligned} \left| \left(\Delta u_{ix} + \Delta \frac{\partial H}{\partial z_{iy}} \right)_{y=b} \right| &\leq S_1\epsilon, \\ \left| \left(\Delta u_i + \Delta \frac{\partial H}{\partial z_{ix}} \right)_{x=a} \right| &\leq S_2\epsilon, \end{aligned}$$

where the constants S_i are independent of ϵ . Consequently, there exists a number T_3 such that $|\eta_3| \leq T_3 \epsilon^2$.

Analogously, from inequality (1.40) and inequalities (1.60) and (1.64) we get

$$|\eta_4| \leq T_4 \epsilon^3.$$

To obtain an estimate of the quantity η_5 in (1.42) we should consider that the functions $\partial H / \partial U_i$ satisfy Lipschitz conditions in the last argument and that the $\partial^2 H / \partial U_i \partial U_j$ are bounded. Then we get

$$|\eta_5| \leq T_5 \epsilon^2.$$

Analogously, we find that the estimates

$$|\eta_i| \leq T_i \epsilon^2, \quad i = 6, \dots, 13,$$

where the constants T_i must be taken independent of ϵ , are valid for the quantities η_6, \dots, η_{13} .

Thus, the quantity η occurring in inequality (1.52) satisfies the inequality

$$(1.68) \quad |\eta| \leq T \epsilon^2, \quad T = \text{const.}$$

By the definition of control v_ϵ and by virtue of inequalities (1.62) and (1.68), we shall have

$$\begin{aligned} & - \iint_G [H(x, y, U(x, y), v_\epsilon) - H(x, y, U(x, y), v)] dy dx - \eta \\ & = \iint_G [H(x, y, U(x, y), v_1) - H(x, y, U(x, y), v)] dy dx - \eta \\ & \leq -\epsilon \delta + T \epsilon^2. \end{aligned}$$

The number ϵ can be chosen so small that the right-hand side of the last inequality becomes negative, but this contradicts inequality (1.52).

Theorem 1 is proved.

II. OPTIMAL PROCESSES IN SYSTEMS WHOSE BEHAVIOR IS DESCRIBED BY PARABOLIC EQUATIONS

2.1. Statement of the problem. Optimality conditions. Let E^n be the Euclidean space of the vectors $x = (x_1, \dots, x_n)$ and let G be a region in this space bounded by a class $A^{(2)}$ surface (see [11, p. 10]). Let $X_i(x)$ denote the direction cosines of the outer normals to boundary Γ .

Further, in the region $G + \Gamma$ we define the elliptic operator $L = (L_1, \dots, L_m)$,

$$(2.1) \quad L_i y = \sum_{\nu=1}^m \sum_{j,k=1}^n a_{jk}^{i\nu} \frac{\partial^2 y_\nu}{\partial x_j \partial x_k},$$

where the functions $a_{jk}^{i\nu}(x)$ are of class $C^{(2)}$ in the region $G + \Gamma$. We let $M = (M_1, \dots, M_m)$ denote the operator defined by the formula

$$M_i z = \sum_{\nu=1}^m \left[\sum_{j,k=1}^n \frac{\partial}{\partial x_j} \left(a_{jk}^{i\nu} \frac{\partial z_\nu}{\partial x_k} \right) + \sum_{j=1}^n \frac{\partial}{\partial x_j} (e_j^{i\nu} z_\nu) \right], \quad e_j^{i\nu} = - \sum_k \frac{\partial a_{jk}^{i\nu}}{\partial x_k}.$$

By direct verification we can convince ourselves of the validity of the formula

$$\begin{aligned} \sum_{i=1}^m \int_G (z_i L_i y - y_i M_i z) dx \\ = \sum_{i,\nu=1}^m \sum_{j=1}^n \int_\Gamma \left[\sum_{k=1}^n a_{jk}^{i\nu} \left(z_i \frac{\partial y_\nu}{\partial x_k} - y_\nu \frac{\partial z_i}{\partial x_k} \right) + e_j^{i\nu} y_\nu z_i \right] X_j(x) d\sigma. \end{aligned}$$

By the same method which is applied for one elliptic equation, we can transform this formula to

$$(2.2) \quad \sum_{i=1}^m \int_G (z_i L_i y - y_i M_i z) dx = \sum_{i=1}^m \int_\Gamma (z_i P_i y - y_i Q_i z) d\sigma,$$

where

$$(2.3) \quad \begin{aligned} P_i y &= \sum_{\nu=1}^m \left[a_{i\nu}^{i\nu} \frac{dy_\nu}{dl_{i\nu}} + b_{i\nu} y_\nu \right], \\ Q_i z &= \sum_{\nu=1}^m \left[a_{\lambda\nu}^{i\nu} \frac{dz_\nu}{d\lambda_{i\nu}} + d_{i\nu} z_\nu \right]. \end{aligned}$$

In (2.3) the directions $l_{i\nu}$ are chosen arbitrarily except that $\cos(n, l_{i\nu}) > 0$ (n is the outer normal to Γ) and their direction cosines are of class $C^{(1)}$ on Γ . The directions $\lambda_{i\nu}$ are chosen independently of $l_{i\nu}$.

Now let the coefficients in the operator L depend also on the variable t , $0 \leq t \leq T$. We shall study controlled processes described by the system of parabolic equations

$$(2.4) \quad L_t y = f(t, x, y_x, u) \quad \left(L_{it} y = \frac{\partial y_i}{\partial t} - L_i y \right),$$

where $0 \leq t \leq T$, $x \in G$, the function $f = (f_1, \dots, f_m)$ is continuous in t and twice continuously differentiable in all the remaining arguments. The parameter u takes values in some bounded (open or closed) region U of a p -dimensional Euclidean space.

Let us assume further that the function $y(t, x) = (y_1, \dots, y_m)$, defined by the system of equations (2.4), also satisfies the condition

$$(2.5) \quad y(0, x) = a(x), \quad x \in G,$$

where $a(x)$ is a continuous vector-valued function. The boundary conditions are chosen in one of the following forms:

$$(2.6_1) \quad y_i(t, x) = \phi_i(t, x), \quad x \in \Gamma, \quad 0 \leq t \leq T,$$

or

$$(2.6_2) \quad P_i(t, x)y = \phi_i(t, x, y, v), \quad x \in \Gamma, \quad 0 \leq t \leq T,$$

where the operators P_i are defined by (2.3) in which the functions $a_i^{is}(t, x)$ and $b_{iv}(t, x)$ are continuously differentiable, while the functions ϕ_i satisfy the same conditions that the functions f_i do. The parameter v takes values in some bounded (open or closed) region V of a q -dimensional Euclidean space.

In what follows we shall speak about the first or the second boundary value problem depending on whether the boundary conditions have been chosen in form (2.6₁) or (2.6₂).

The function $u(t, x)$ will be called an admissible control in the first boundary value problem if all its components are piecewise-continuous and it takes its values in the control region U . The surfaces of discontinuity of admissible controls are assumed to be smooth and each of them either is orthogonal to the t -axis or is such that in the neighborhood of any point of it we can introduce a nonsingular coordinate transformation, $\tau = t$, $\xi_i = \xi_i(t, x)$, $i = 1, \dots, n$, such that the surface of discontinuity becomes a piece of the plane $\xi_n = 0$. Just as in Part I, we do not exclude the possibility here that the form of the dependency of the individual components of the vector $u(t, x)$ on x and t is given in advance. In particular, they may depend only on t or only on x .

If the discontinuities of some admissible control satisfy the first of the stated conditions, then the first boundary value problem (2.4)–(2.6) corresponding to this control splits up into several problems of the same kind but in regions which abut each other along the surfaces of discontinuity of the control. In this case the boundary value problem has a unique continuous solution (for example, see [2]) which, moreover, is not subject to any supplementary smoothness conditions on the surfaces of discontinuity of the control.

However, if these surfaces are not planes orthogonal to the t -axis, then by a solution of the first boundary value problem we shall mean a vector-valued function $y(t, x)$ which satisfies the system of equations (2.4), the conditions (2.5) and (2.6₁), and also certain smoothness conditions on the surfaces of discontinuity of the control. Apparently this problem has not been studied in its general form; however, special cases of it have been considered in a number of papers (for example, see [13], [14], [15]), where various existence and uniqueness theorems have been obtained. Therefore, in what follows we shall assume that the given functions in (2.4) and in conditions (2.5) and (2.6₁) possess not only the properties listed above but

also satisfy the further condition that a unique solution of the first boundary value problem corresponds to every admissible control.

In the second boundary value problem as an admissible control we shall take a function $\omega(t, x) = \{u(t, x), v(t, x)\}$, where $u(t, x)$ satisfies the same conditions as does the admissible control in the first boundary value problem, while $v(t, x)$ is a piecewise-continuous function with values in region V . The discontinuities of function $v(t, x)$ should satisfy the same conditions that the discontinuities of function $u(t, x)$ do. In what follows, furthermore, we shall assume that to every admissible control $\omega(t, x)$ there corresponds a unique solution of the second boundary value problem.

On the set of admissible controls we define the functionals

$$(2.7_1) \quad S_1 = \sum_{i=1}^m \left[\int_G \alpha_i(x) y_i(T, x) dx + \int_0^T \int_G \beta_i(t, x) y_i(t, x) dx dt + \int_0^T \int_\Gamma \gamma_i(t, x) P_i(t, x) y d\sigma dt \right],$$

$$(2.7_2) \quad S_2 = \sum_{i=1}^m \left[\int_G \alpha_i(x) y_i(T, x) dx + \int_0^T \int_G \beta_i(t, x) y_i(t, x) dx dt + \int_0^T \int_\Gamma \gamma_i(t, x) y_i(t, x) d\sigma dt \right],$$

where α_i , β_i and γ_i are given continuous functions, and the functions $y_i(t, x)$ in functional S_1 form the solution of the first boundary value problem, while in functional S_2 , of the second boundary value problem. The quantity T is not fixed and can change with the transition from one control to another.

We shall say that an admissible control $u(t, x)$ ($\omega(t, x)$), $0 \leq t \leq T$, in the first (second) boundary value problem transfers the system from the state (2.5) to the state (2.8) if the corresponding solution of this problem, satisfying condition (2.5) at $t = 0$, at the instant $t = T$ satisfies the conditions

$$(2.8) \quad \Phi_\alpha(T, x, y(T, x)) = 0, \quad \int_G \Psi_\beta(T, x, y(T, x), y_x(T, x)) dx = c_\beta, \\ \alpha = 1, \dots, j, \quad \beta = 1, \dots, k,$$

where $y(T, x) = \{y_1(T, x), \dots, y_m(T, x)\}$, c_β are given constants, and $j + k \leq m$.

Let $u(t, x)$, $0 \leq t \leq T$, $x \in G$, be some admissible control in the first boundary value problem (2.4)–(2.6), transferring the system from the state (2.5) to the state (2.8). Just as in the case of hyperbolic equations, here too we should require that the class of admissible controls be complete

in order that, first, the optimal problem not be trivial and, second, it be solvable by the method presented above.

In what follows we shall assume that the completeness of the class of admissible controls is defined by the following fundamental properties, formulated here for the first boundary value problem.

Let $u(t, x)$ and $u_1(t, x)$ be two admissible controls in the first boundary value problem (2.4)–(2.6), defined in the regions C , $0 \leq t \leq T$, $x \in G$, and C_1 , $0 \leq t \leq T_1$, $x \in G$, and transferring the system from the state (2.5) to the state (2.8). Then, for an arbitrarily small positive number ϵ we can find a control $u_\epsilon(t, x)$ defined in the region C_ϵ , $0 \leq t \leq T_\epsilon$, $x \in G$, such that:

(i) it transfers the system from the state (2.5) to the state (2.8) and, moreover, condition (2.8) is fulfilled when $t = T_\epsilon$;

(ii) it is defined in the region $C_\epsilon \times C$ by the formula

$$u_\epsilon(t, x) = \begin{cases} u_1 & \text{when } (t, x) \in G_\epsilon, \\ u & \text{when } (t, x) \in C_\epsilon \times C \setminus G_\epsilon, \end{cases}$$

where G_ϵ is an arbitrary given region (whose area equals ϵ) lying strictly inside C ;

(iii) the inequality $|T - T_\epsilon| \leq L\epsilon$ is satisfied, where the constant L is independent of ϵ .

The condition for the completeness of the class of admissible controls in the second boundary value problem (2.4)–(2.6) is defined analogously.

We pose the problem: from among the admissible controls in the first (second) boundary value problem which transfer the system from the state (2.5) to the state (2.8), to find the control such that the corresponding solution of this problem would realize the minimum of functional S_1 (S_2).

The admissible control and the corresponding solution of the optimal problem being considered will be called the control and the solution optimal relative to S_1 (S_2).

Optimal control problems in processes described by different boundary value problems for parabolic equations are of theoretical and practical interest. A number of papers [7], [16], [17] have treated certain problems in the case when the control is realized by initial or boundary conditions. As the optimality criterion, either time optimality is chosen or the functional

$I = \int_0^1 [u(T, x) - u_0(x)]^2 dx + \gamma \int_0^T p^2(t) dt$, where $u_0(x)$ is a given function in $L_2(0, 1)$, $p(t)$ is the control, and γ is a nonnegative constant.

In this article we shall consider the problem when the control in the process can be effected simultaneously by controls which occur in the system equations and in the boundary conditions. Here, at first, we consider the

problem where as the optimality criteria we select the functionals S_1 and S_2 . The general cases will be treated at the end of Part II. The time-optimal problem also will be treated there.

In order to formulate the optimality conditions we introduce the auxiliary functions H and h by setting

$$H(t, x, w, u) = \sum_{i=1}^m z_i f_i(t, x, y, y_x, u), \quad h(t, x, p, v) = \sum_{i=1}^m z_i \phi_i(t, x, y, v),$$

where

$$p = (z_1, \dots, z_m, y_1, \dots, y_m),$$

$$w = \left(z_1, \dots, z_m, y_1, \dots, y_m, \frac{\partial y_1}{\partial x_1}, \dots, \frac{\partial y_m}{\partial x_n} \right).$$

To determine the functions $z_i = z_i(t, x)$ we take some admissible control $u(t, x)$ ($\omega(t, x) = \{u(t, x), v(t, x)\}$) in the first (second) boundary value problem (2.4)–(2.6). To it we correspond the solution $y = y(t, x)$ of this problem. We introduce the functions $z_i = z_i(t, x)$ by means of the differential equations

$$(2.9) \quad M_{it} z = - \frac{\partial H(t, x, w, u)}{\partial y_i} + \sum_{k=1}^n \frac{d}{dx_k} \left(\frac{\partial H(t, x, w, u)}{\partial y_{ik}} \right) + \beta_i(t, x),$$

$$y_{ik} = \frac{\partial y_i}{\partial x_k}, \quad 0 \leq t \leq T, \quad x \in G,$$

with initial conditions

$$(2.10) \quad z_i(T, x) = -\alpha_i(x) - \sum_{\alpha=1}^j a_\alpha(x) \frac{\partial \Phi_\alpha}{\partial y_i} - \sum_{\beta=1}^k b_\beta \left(\frac{\partial \Psi_\beta}{\partial y_i} - \sum_{\nu=1}^n \frac{d}{dx_\nu} \left[\frac{\partial \Psi_\beta}{\partial y_{i\nu}} \right] \right), \quad x \in G,$$

where $M_{it} z = \partial z_i / \partial t + M_i z$, the functions $\alpha_i(x)$, $\beta_i(t, x)$ are taken from the functionals S_1 and S_2 , the functions Φ_α and Ψ_β are taken from conditions (2.8), while the constants b_β and the functions $a_\alpha(x)$ are as yet undefined. The choice of the boundary conditions for (2.9) depends upon whether we are considering the first or the second boundary value problem (2.4)–(2.6).

If it is the first boundary value problem that we are considering, the boundary conditions are chosen in the form

$$(2.11_1) \quad z_i(t, x) = \gamma_i(t, x), \quad x \in \Gamma, \quad i = 1, \dots, m,$$

where the $\gamma_i(t, x)$ are the functions occurring in the definition of functional

S_1 . If the second boundary value problem (2.4)–(2.6) is being considered, the boundary conditions for (2.9) should be chosen as

$$(2.11_2) \quad Q_i z = \frac{\partial h(t, x, p, v)}{\partial y_i} + \sum_{k=1}^n \frac{\partial H(t, x, w, u)}{\partial y_{ik}} X_k(x) - \gamma_i(t, x),$$

$$x \in \Gamma, i = 1, \dots, m,$$

where the Q_i are the operators introduced by formulas (2.3), $X_k(x)$ are the direction cosines of the outer normals to the surface Γ , $\gamma_i(t, x)$ are the functions occurring in the definition of functional S_2 .

Both the boundary value problems by means of which we have introduced the functions $z_i(t, x)$ are linear and satisfy the same conditions as do the corresponding (first or second) boundary value problem (2.4)–(2.6). Furthermore, it is natural to require that the consistency condition be fulfilled at $t = T$.

Therefore, we can take it that for given functions $a_\alpha(x)$ and given constants b_β , each admissible control uniquely determines a solution of the corresponding auxiliary boundary value problem.

Using (2.2) we can show that for any twice piecewise-continuously differentiable functions $y_i(t, x)$ and $z_i(t, x)$, $i = 1, \dots, m$, the Green-Ostrogradskii formula

$$\begin{aligned} \sum_{i=1}^m \int_0^T \int_G (z_i L_{ii} y + y_i M_{ii} z) dx dt \\ = \sum_{i=1}^m \left[\int_0^T \int_\Gamma (y_i Q_i z - z_i P_i y) d\sigma dt + \int_G y_i z_i \Big|_{t=0}^T dx \right], \end{aligned}$$

which relates the operators occurring in the formulation of the original and of the auxiliary boundary value problems, is valid.

Let $u(t, x)$ ($\omega(t, x) = \{u(t, x), v(t, x)\}$) be some admissible control in the first (second) boundary value problem, transferring the system from the state (2.5) to the state (2.8), and let $y(t, x)$ and $z(t, x)$ be the corresponding solutions of the first (second) boundary value problems (2.4)–(2.6) and (2.9)–(2.11). We introduce the functionals

$$(2.13) \quad J_1[u] = \int_0^T \int_G H(t, x, w(t, x), u) dx dt,$$

$$(2.14) \quad J_2[u] = \int_0^T \int_\Gamma h(t, x, p(t, x), v) d\sigma dt,$$

defined on the piecewise-continuous functions $u(t, x)$ and $v(t, x)$ with values in regions U and V , respectively.

We shall say that the admissible control $\omega(t, x)$, transferring the system from the state (2.5) to the state (2.8), satisfies a maximum condition

relative to the function $z(t, x)$ if for any other control $\omega'(u', v')$, also transferring the system from the state (2.5) to the state (2.8), the inequalities

$$(2.15_1) \quad \Delta J_1[u] = \int_D [H(t, x, w(t, x), u') - H(t, x, w(t, x), u(t, x))] dx dt \leq 0,$$

$$(2.15_2) \quad \Delta J_2[u] = \int_0^\tau \int_\Gamma [h(t, x, p(t, x), v') - h(t, x, p(t, x), v(t, x))] d\sigma dt \leq 0$$

are satisfied, where $D, 0 \leq t \leq \tau, x \in G + \Gamma$, denotes the region in which both the controls ω and ω' are defined.

Analogously we can define the maximum condition for the admissible control $u(t, x)$ in the first boundary value problem (2.4)–(2.6). In this case, only the condition (2.15₁) has to be satisfied.

If we do not impose any restrictions on the form of the dependency of the controls on the arguments x and t , then inequalities (2.15₁) and (2.15₂) are equivalent to the following conditions:

$$H(t, x, w(t, x), u(t, x)) ((=)) \sup_{u \in U} H(t, x, w(t, x), u), \quad x \in G, 0 \leq t \leq T,$$

$$h(t, x, p(t, x), v(t, x)) (=) \sup_{v \in V} h(t, x, p(t, x), v), \quad x \in \Gamma, 0 \leq t \leq T,$$

where the symbol $((=))$ means equality valid everywhere in the region $G, 0 \leq t \leq T, x \in G$, excepting perhaps points lying on a finite number of n -dimensional surfaces whose $(n+1)$ -dimensional volume equals zero. The symbol $(=)$ is defined analogously except that we take $n-1$ and Γ instead of n and G , respectively.

The proof of this assertion is the same as that for the corresponding assertion in Part I.

The necessary conditions for the optimality of the controls in the boundary value problems being considered are given by the following theorems.

THEOREM 5. *In order that an admissible control $u(t, x)$ in the first boundary value problem (2.4)–(2.6), transferring the system from the state (2.5) to the state (2.8), and the corresponding solution $y(t, x)$, be optimal relative to functional S_1 , it is necessary that there exist functions $z_i(t, x)$ and $a_\alpha(x)$ and constants b_β such that:*

(i) *the functions $y(t, x), z(t, x), u(t, x), a(x)$, and the constants b_β form a solution of (2.4) and of (2.9) with the supplementary conditions (2.5), (2.6₁) and (2.10), (2.11₁);*

(ii) *the control $u(t, x)$ satisfies the maximum condition relative to the functions $z(t, x)$;*

(iii) *the condition*

$$\begin{aligned}
 (2.16) \quad & \int_G \left[\sum_{i=1}^m \alpha_i(x) \frac{\partial y_i(T, x)}{\partial T} + \beta_i(T, x) y_i(T, x) \right. \\
 & \left. + \sum_{\alpha=1}^j a_\alpha(x) \frac{d\Phi_\alpha}{dT} + \sum_{\beta=1}^k b_\beta \frac{d\Psi_\beta}{dT} \right] dx \\
 & + \sum_{i=1}^m \int_\Gamma \gamma_i(T, x) P_i(T, x) y_i(T, x) d\sigma = 0
 \end{aligned}$$

is satisfied at the terminal instant $t = T$.

THEOREM 6. *In order that an admissible control $\omega(t, x) = \{u(t, x), v(t, x)\}$ in the second boundary value problem (2.4)–(2.6), transferring the system from the state (2.5) to the state (2.8), and the corresponding solution $y(t, x)$ of this problem be optimal relative to functional S_2 , it is necessary that there exist functions $z_i(t, x)$ and $a_\alpha(x)$ and constants b_β such that:*

- (i) *the functions $y(t, x)$, $z(t, x)$, $\omega(t, x)$, $a(x)$, and the constants b_β form the solution of (2.5), (2.6₂) and (2.10), (2.11₂);*
- (ii) *the control $u(t, x)$ satisfies the maximum condition relative to the functions $z(t, x)$;*
- (iii) *the condition*

$$\begin{aligned}
 (2.17) \quad & \int_G \left[\sum_{i=1}^m \alpha_i(x) \frac{\partial y_i(T, x)}{\partial T} + \beta_i(T, x) y_i(T, x) \right. \\
 & \left. + \sum_{\alpha=1}^j a_\alpha(x) \frac{d\Phi_\alpha}{dT} + \sum_{\beta=1}^k b_\beta \frac{d\Psi_\beta}{dT} \right] dx \\
 & + \sum_{i=1}^m \int_\Gamma \gamma_i(T, x) y_i(T, x) d\sigma = 0
 \end{aligned}$$

is satisfied at the terminal instant $t = T$.

Each of these theorems yields a “complete” system of relations by means of which we can pick out, in general, the isolated controls and their corresponding functions $y(t, x)$, which may be optimal. We convince ourselves of this by the same arguments used in Part I. If in the problem under study it turns out that there are only a finite number of such controls and if an optimal control exists from the physical sense of the problem, then this control can be found by successively comparing all the controls picked out.

We require certain auxiliary relations in order to prove Theorems 5 and 6, which we proceed to derive below.

2.2. Formulas for the increments of functionals J_1 and J_2 . We first consider the first boundary value problem (2.4)–(2.6). On the set of admissible controls we define the functional

$$I_1[u] = \int_G \left[\sum_{i=1}^m z_i L_{it} y - H(t, x, w(t, x), u) \right] dx dt + S_1$$

$$(2.18) \quad + \sum_{\alpha=1}^j \int_G a_{\alpha}(x) \Phi_{\alpha}(T, x, y(T, x)) dx \\ + \sum_{\beta=1}^k b_{\beta} \left[\int \Psi_{\beta}(T, x, y(T, x), y_x(T, x)) dx - C_{\beta} \right],$$

in which the functional S_1 plays the role of the optimality criterion, the functions Φ_{α} and Ψ_{β} are taken from relations (2.8), and the functions $y(t, x)$, $z(t, x)$, $a(x)$ and the constants b_{β} are considered given.

Let $u(t, x)$, $0 \leq t \leq T$, $x \in G$, be some admissible control transferring the system from the state (2.5) to the state (2.8), and let $y(t, x)$ be the corresponding solution of the first boundary value problem (2.4)–(2.6). Then $I_1[u(t, x)] = S_1$ for arbitrary functions $z_i(t, x)$, $a_{\alpha}(x)$ and constants b_{β} . Hence, if to every admissible control $u(t, x)$ which transfers the system from the state (2.5) to the state (2.8) we associate a solution of the first boundary value problem (2.4)–(2.6) when $u = u(t, x)$, then at these controls the functional I_1 will achieve its own extremal value simultaneously with S_1 independently of the choice of the functions $z(t, x)$, $a(x)$ and the constants b_{β} .

We take an admissible control $u_1(t, x)$, $0 \leq t \leq T_1$, $x \in G$, different from u , which also transfers the system from the state (2.5) to the state (2.8) at the instant $t = T_1$. Let D , $0 \leq t \leq \tau = \min \{T, T_1\}$, $x \in G$, denote the intersection of the regions C , $0 \leq t \leq T$, $x \in G$, and C_1 , $0 \leq t \leq T_1$, $x \in G$. Both of the controls $u(t, x)$ and $u_1(t, x)$, and also their corresponding functions $y(t, x, u)$, $z(t, x, u)$ and $y(t, x, u_1)$, $z(t, x, u_1)$, which form the solutions of (2.4) and (2.9) with supplementary conditions (2.5), (2.6₁) and (2.10), (2.11₁), are defined in region D . Hence it follows that the functions $\Delta y_i = y_i(t, x, u_1) - y_i(t, x, u)$ and $\Delta z_i = z_i(t, x, u_1) - z_i(t, x, u)$ satisfy the equations

$$(2.19) \quad L_{it} \Delta y = \Delta \frac{\partial H}{\partial z_i}, \quad M_{it} \Delta z = -\Delta \frac{\partial H}{\partial y_i} + \sum_{k=1}^n \frac{d}{dx_k} \left(\Delta \frac{\partial H}{\partial y_{ik}} \right)$$

and the supplementary conditions

$$(2.20) \quad \Delta y_i(0, x) = 0, \quad x \in G,$$

$$(2.21) \quad \Delta y_i(t, x) = \Delta z_i(t, x) = 0, \quad x \in \Gamma,$$

where

$$\Delta \frac{\partial H}{\partial w_i} = \frac{\partial H(t, x, w(t, x, u_1), u_1)}{\partial w_i} - \frac{\partial H(t, x, w(t, x, u), u)}{\partial w_i}.$$

If the control $u(t, x)$ is optimal relative to functional S_1 , then in ac-

cordance with what we have said above the following inequality will be valid.

$$(2.22) \quad \Delta I_1[u] = I_1[u_1] - I_1[u] \geq 0.$$

We transform the increments of the functionals occurring in this inequality, taking, for definiteness, that $\tau = T \leq T_1$, i.e., $D = C$. The auxiliary functions $a_\alpha(x)$ and the constants b_β are taken to be arbitrary but fixed.

We denote the first integral in functional (2.18) by $K_1[u]$. Then we have

$$(2.23) \quad \Delta K_1[u] = \int_C \left[\sum_{i=1}^m (\Delta z_i L_{it} \Delta y + \Delta z_i L_{it} y + z_i L_{it} \Delta y) - \Delta H \right] dx dt,$$

where $\Delta w_i = w_i(t, x, u_1) - w_i(t, x, u)$. During the computation the increment ΔK_1 of the integral over the region $C_1 \setminus C$ is omitted since its integrand is identically zero.

The functions Δy_i and Δz_i satisfy (2.19) as well as conditions (2.20) and (2.21). Therefore, applying the Green-Ostrogradskii formula we get

$$\begin{aligned} \int_C \sum_{i=1}^m \Delta z_i L_{it} \Delta y dx dt &= \int_C \sum_{i=1}^m \left[\Delta \frac{\partial H}{\partial y_i} \Delta y_i + \sum_{k=1}^n \Delta \frac{\partial H}{\partial y_{ik}} \Delta y_{ik} \right] dx dt \\ &\quad + \int_G \sum_{i=1}^m \Delta y_i(T, x) \Delta z_i(T, x) dx. \end{aligned}$$

On the other hand

$$\int_C \sum_{i=1}^m \Delta z_i L_{it} \Delta y dx dt = \int_C \sum_{i=1}^m \Delta \frac{\partial H}{\partial z_i} \Delta z_i dx dt.$$

From the last two equations it follows that

$$(2.24) \quad \begin{aligned} &\int_C \sum_{i=1}^m \Delta z_i L_{it} \Delta y dx dt \\ &= \frac{1}{2} \left\{ \int_C \sum_{i=1}^N \Delta \frac{\partial H}{\partial w_i} \Delta w_i dx dt + \int_G \sum_{i=1}^m \Delta y_i(T, x) \Delta z_i(T, x) dx \right\}, \end{aligned}$$

where $N = (n + 2)m$ is the dimension of vector $w_i = (z_1, \dots, z_m, y_1, \dots, y_m, y_{11}, \dots, y_{mn})$.

Analogously, from (2.4) we obtain

$$(2.25) \quad \int_C \sum_{i=1}^m \Delta z_i L_{it} y dx dt = \int_C \sum_{i=1}^m \frac{\partial H}{\partial z_i} \Delta z_i dx dt.$$

Since the functions Δy_i satisfy (2.19) and the supplementary conditions (2.20), (2.21), and since the z_i form the solutions of the first boundary

value problem (2.9)–(2.11), we have

$$(2.26) \quad \int_C \sum_{i=1}^m z_i L_{it} \Delta y \, dx \, dt = \Delta \bar{S}_1 + \int_C \sum_{i=1}^m \left[\frac{\partial H}{\partial y_i} \Delta_u y_i + \sum_{k=1}^n \frac{\partial H}{\partial y_{ik}} \Delta y_{ik} \right] dx \, dt,$$

where

$$\begin{aligned} \Delta_u y(t, x, u) &= y(t, x, u_1) - y(t, x, u), \\ \Delta \bar{S}_1 &= \sum_{i=1}^m \left\{ \int_G z_i(T, x) \Delta_u y_i(T, x, u) \, dx - \int_C \beta_i(t, x) \Delta_u y_i(t, x, u) \, dx \, dt \right. \\ &\quad \left. - \int_0^T \int_\Gamma z_i(t, x) P_i \Delta_u y(t, x, u) \, d\sigma \, dt \right\}. \end{aligned}$$

From relations (2.23)–(2.26) we get

$$\begin{aligned} \Delta K_1[u] &= \Delta \bar{S}_1 + \int_C \left[\sum_{i=1}^N \left(\frac{1}{2} \Delta \frac{\partial H}{\partial w_i} + \frac{\partial H}{\partial w_i} \right) \Delta w_i - \Delta H \right] dx \, dt \\ &\quad + \frac{1}{2} \int_G \sum_{i=1}^m \Delta_u y_i(T, x) \Delta_u z_i(T, x) \, dx. \end{aligned}$$

Applying Taylor's formula to the functions H and $\partial H / \partial w_i$ and restricting ourselves to second order terms in the expansions of these functions, we get

$$(2.27) \quad \Delta K_1[u] = \Delta \bar{S}_1 - \int_C [H(t, x, w(t, x), u_1) - H(t, x, w(t, x), u)] \, dx \, dt - \eta_1 - \eta_2$$

where

$$\begin{aligned} (2.28) \quad \eta_1 &= \frac{1}{2} \sum_{i=1}^N \int_G \left[\frac{\partial H(t, x, w(t, x, u), u_1)}{\partial w_i} - \frac{\partial H(t, x, w(t, x, u), u)}{\partial w_i} \right. \\ &\quad \left. + \sum_{k=1}^m \left(\frac{\partial^2 H(t, x, w(t, x, u) + \theta_1 \Delta w, u_1)}{\partial w_i \partial w_k} \right. \right. \\ &\quad \left. \left. - \frac{\partial^2 H(t, x, w(t, x, u) + \theta_2 \Delta w, u)}{\partial w_i \partial w_k} \right) \Delta w_k \right] \Delta w_i \, dx \, dt, \\ \eta_2 &= \frac{1}{2} \int_G \sum_{i=1}^m \Delta_u y_i(T, x) \Delta_u z_i(T, x) \, dx, \quad 0 \leq \theta_i \leq 1. \end{aligned}$$

If we compute the increment of functional S_1 by the formula of finite increments, we have

$$\Delta S_1 = \sum_{i=1}^m \left\{ \int_G \left[\alpha_i(x) \frac{\partial y_i(T, x, u)}{\partial T} + \beta_i(T, x) y_i(T, x, u) \right] dx \right.$$

$$\begin{aligned}
(2.29) \quad & + \int_{\Gamma} \gamma_i(T, x) P_i(T, x) y_i(T, x, u) d\sigma \Big\} \Delta T \\
& + \sum_{i=1}^m \left\{ \int_G \alpha_i(x) \Delta_u y_i(T, x, u) dx \right. \\
& + \int_0^T \int_{\Gamma} \beta_i(t, x) \Delta_u y(t, x, u) dx dt \\
& \left. + \int_0^T \int_{\Gamma} \gamma_i(t, x) P_i(t, x) \Delta_u y(t, x, u) d\sigma dt \right\} - \eta_3,
\end{aligned}$$

where

$$\begin{aligned}
(2.30) \quad & \Delta_u y(t, x, u) = y_i(t, x, u_1) - y_i(t, x, u), \\
& \eta_3 = \sum_{i=1}^m \left\{ \int_G \left(\alpha_i(x) \sum_{j=1}^p \left[\frac{\partial y_i(T, x, u)}{\partial T} - \frac{\partial y_i(\bar{t}_j, x, u_1)}{\partial T} \right] \Delta t_j \right. \right. \\
& + [\beta_i(T, x) y_i(T, x, u) - \beta_i(\bar{t}, x) y_i(\bar{t}, x, u_1)] \Delta T \Big) dx \\
& + \int_{\Gamma} [\gamma_i(T, x) P_i(T, x) y_i(T, x, u) \\
& \left. - \gamma_i(\bar{t}, x) P_i(\bar{t}, x) y_i(\bar{t}, x, u_1)] \Delta T d\sigma \right\}.
\end{aligned}$$

Here the bar over the argument t signifies that we have chosen a certain value of it from the interval $[T, T_1]$, while $t_j = t_j(x)$ is the surface of discontinuity of control u_i in the region $0 \leq t \leq T_1, x \in G$.

Analogously (recalling that the symbols d/dT and d/dx_k denote the total derivatives with respect to T and x_k , respectively) we find that

$$\begin{aligned}
(2.31) \quad & \sum_{\alpha=1}^j \int_G a_{\alpha}(x) [\Phi_{\alpha}(T_1, x, y(T_1, x, u_1)) - \Phi_{\alpha}(T, x, y(T, x, u))] dx \\
& = \sum_{\alpha=1}^j \int_G a_{\alpha}(x) \left[\frac{d\Phi_{\alpha}(T, x, y(T, x, u))}{dT} \Delta T \right. \\
& \left. + \sum_{i=1}^m \frac{\partial \Phi_{\alpha}(T, x, y(T, x, u))}{\partial y_i} \Delta_u y(T, x, u) \right] dx - \eta_4 - \eta_5, \\
& \sum_{\beta=1}^k b_{\beta} \int_G [\Psi_{\beta}(T_1, x, y(T_1, x, u_1), y_x(T_1, x, u_1)) \\
& - \Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))] dx \\
& = \sum_{\beta=1}^k b_{\beta} \int_G \left\{ \frac{d\Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{dT} \Delta T \right.
\end{aligned}$$

$$+ \sum_{i=1}^m \left[\frac{\partial \Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{\partial y_i} \right. \\ \left. - \sum_{k=1}^n \frac{d}{dx_k} \left(\frac{\partial \Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{\partial y_{ik}} \right) \right] \Delta_u y_i(T, x, u) \Big\} dx - \eta_6 - \eta_7,$$

where

$$\begin{aligned} \eta_4 &= \sum_{\alpha=1}^j \int_G a_{\alpha}(x) \sum_{i=1}^p \left[\frac{d\Phi_{\alpha}(T, x, y(T, x, u))}{dT} \right. \\ &\quad \left. - \frac{d\Phi_{\alpha}(\bar{t}_i, x, y(\bar{t}_i, x, u_1))}{dT} \right] \Delta t_i dx, \\ \eta_5 &= \sum_{\alpha=1}^j \sum_{i=1}^m \int_G a_{\alpha}(x) \left[\frac{\partial \Phi_{\alpha}(T, x, y(T, x, u))}{\partial y_i} - \frac{\partial \Phi_{\alpha}(T, x, \bar{y})}{\partial y_i} \right] \Delta_u y(T, x, u) dx, \\ \eta_6 &= \sum_{\beta=1}^k b_{\beta} \int_G \sum_{i=1}^p \left[\frac{d\Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{dT} \right. \\ &\quad \left. - \frac{d\Psi_{\beta}(\bar{t}_i, x, y(\bar{t}_i, x, u_1), y_x(\bar{t}_i, x, u_1))}{dT} \right] \Delta t_i dx, \\ \eta_7 &= \sum_{\beta=1}^k b_{\beta} \int_G \sum_{i=1}^m \left[\left(\frac{\partial \Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{\partial y_i} \right. \right. \\ &\quad \left. \left. - \frac{\partial \Psi_{\beta}(T, x, \bar{y}, \bar{y}_x)}{\partial y_i} \right) \Delta_u y_i(T, x, u) \right. \\ &\quad \left. + \sum_{k=1}^n \left(\frac{\partial \Psi_{\beta}(T, x, y(T, x, u), y_x(T, x, u))}{\partial y_{ik}} \right. \right. \\ &\quad \left. \left. - \frac{\partial \Psi_{\beta}(T, x, \bar{y}, \bar{y}_x)}{\partial y_{ik}} \right) \Delta_u y_{ik}(T, x, u) \right] dx. \end{aligned}$$

Since the functions $z_i(t, x)$, $a_{\alpha}(x)$ and the constants b_{β} satisfy conditions (2.10), (2.11₁) and (2.16), it follows from inequality (2.22) by virtue of (2.27), (2.29) and (2.31) that

$$(2.32) \quad -\Delta J_1 = -\int_G [H(t, x, w(t, x, u), u_1) - H(t, x, w(t, x, u), u)] dx dt \geq \eta,$$

where

$$\eta = \sum_{i=1}^7 \eta_i.$$

Formula (2.32) gives an estimate of the increment of functional J_1 for a transition from control u to control u_1 .

We now consider the problem of minimizing functional S_2 in the second

boundary value problem (2.4)–(2.6). To obtain the formula for the increment of functional J_2 defined by (2.14) we consider the auxiliary functional

$$\begin{aligned}
 I_2[\omega] = & \int_C \left[\sum_{i=1}^m z_i L_{it} y - H(t, x, w(t, x, u), u) \right] dx dt \\
 & + \int_0^T \int_\Gamma \left[\sum_{i=1}^m z_i P_i y - h(t, x, p(t, x), v) \right] d\sigma dt \\
 (2.33) \quad & + S_2 + \sum_{\alpha=1}^j \int_G a_\alpha(x) \Phi_\alpha(T, x, y(T, x)) dx \\
 & + \sum_{\beta=1}^k b_\beta \left[\int_G \Psi_\beta(T, x, y(T, x), y_x(T, x)) dx - c_\beta \right],
 \end{aligned}$$

where the operators P_i are defined in (2.3), while the functional S_2 plays the role of the optimality criterion in the optimal problem being considered.

Just as in the case of the first boundary value problem (2.4)–(2.6), we can show that if the control $\omega(t, x) = \{u(t, x), v(t, x)\}$, transferring the system from the state (2.5) to the state (2.8), is optimal relative to functional S_2 , then for any other control ω_1 , also transferring the system from the state (2.5) to the state (2.8), the inequality

$$(2.34) \quad \Delta I_2[\omega(t, x)] = I_2[\omega_1] - I_2[\omega] \geq 0$$

is valid, where $y(t, x)$ is the solution of the second boundary value problem (2.4)–(2.6) corresponding to the control $\omega(t, x)$, and $z(t, x)$ is the solution of the second boundary value problem (2.9)–(2.11) corresponding to the same control. The functions $a_\alpha(x)$ and the constants b_β are considered as given.

Let the optimal control $\omega(t, x)$ be defined in the region C , $0 \leq t \leq T$, $x \in G + \Gamma$, and the control $\omega_1(t, x) = \{u_1(t, x), v_1(t, x)\}$ in the region C_1 , $0 \leq t \leq T_1$, $x \in G + \Gamma$. Let D , $0 \leq t \leq \tau = \min \{T, T_1\}$, $x \in G + \Gamma$, denote the intersections of regions C and C_1 . Both of the controls $\omega(t, x)$ and $\omega_1(t, x)$, and also their corresponding functions $y(t, x, \omega)$, $z(t, x, \omega)$ and $y(t, x, \omega_1)$, $z(t, x, \omega_1)$, forming the solutions of (2.4) and (2.9) with supplementary conditions (2.5), (2.6₂) and (2.10), (2.11₂), are defined in the region D . Hence it follows that the functions $\Delta y_i = y_i(t, x, \omega_1) - y_i(t, x, \omega)$, $\Delta z_i = z_i(t, x, \omega_1) - z_i(t, x, \omega)$ satisfy (2.19) and also condition (2.20) and the conditions

$$\begin{aligned}
 (2.35) \quad P_i \Delta y &= \Delta \frac{\partial h}{\partial z_i}, \quad Q_i \Delta z = \Delta \frac{\partial h}{\partial y_i} + \sum_{k=1}^n \Delta \frac{\partial H(t, x, w, u)}{\partial y_{ik}} X_k(x), \\
 &x \in \Gamma, \quad 0 \leq t \leq \tau, \quad i = 1, \dots, m.
 \end{aligned}$$

We transform the increments of all the functionals occurring in inequality (2.34), taking, for definiteness, that $T \leq T_1$.

We denote the sum of the first two integrals in functional (2.33) by $K_2[\omega]$ and we have

$$(2.36) \quad \begin{aligned} \Delta K_2[\omega] = & \int_G \left[\sum_{i=1}^m (\Delta z_i L_{it} \Delta y + \Delta z_i L_{it} y + z_i L_{it} \Delta y) - \Delta H \right] dx dt \\ & + \int_0^T \int_\Gamma \left[\sum_{i=1}^m (\Delta z_i P_i \Delta y + \Delta z_i P_i y + z_i P_i \Delta y) - \Delta h \right] d\sigma dt. \end{aligned}$$

Since the functions Δy_i and Δz_i satisfy (2.19) with supplementary conditions (2.20) and (2.35), by applying the Green-Ostrogradskii formula we get

$$\begin{aligned} & \sum_{i=1}^m \left[\int_G \Delta z_i L_{it} \Delta y dx dt + \int_0^T \int_\Gamma \Delta z_i P_i \Delta y d\sigma dt \right] \\ & = \sum_{i=1}^m \left[\int_G \left(\Delta \frac{\partial H}{\partial y_i} \Delta y_i + \sum_{k=1}^n \Delta \frac{\partial H}{\partial y_{ik}} \Delta y_{ik} \right) dx dt \right. \\ & \quad \left. + \int_0^T \int_\Gamma \Delta \frac{\partial h}{\partial y_i} \Delta y_i d\sigma dt + \int_G \Delta y_i(T, x) \Delta z_i(T, x) dx \right]. \end{aligned}$$

On the other hand,

$$\begin{aligned} & \sum_{i=1}^m \left[\int_G \Delta z_i L_{it} \Delta y dx dt + \int_0^T \int_\Gamma \Delta z_i P_i \Delta y d\sigma dt \right] \\ & = \sum_{i=1}^m \left[\int_G \Delta \frac{\partial H}{\partial z_i} \Delta z_i dx dt + \int_0^T \int_\Gamma \Delta \frac{\partial h}{\partial z_i} \Delta z_i d\sigma dt \right]. \end{aligned}$$

From these formulas it follows that

$$(2.37) \quad \begin{aligned} & \sum_{i=1}^m \left[\int_G \Delta z_i L_{it} \Delta y dx dt + \int_0^T \int_\Gamma \Delta z_i P_i \Delta y d\sigma dt \right] \\ & = \frac{1}{2} \left[\sum_{i=1}^N \int_G \Delta \frac{\partial H}{\partial w_i} \Delta w_i dx dt + \sum_{i=1}^{2m} \int_0^T \int_\Gamma \Delta \frac{\partial h}{\partial p_i} \Delta p_i d\sigma dt \right. \\ & \quad \left. + \sum_{i=1}^m \int_G \Delta y_i(T, x) \Delta z_i(T, x) dx \right]. \end{aligned}$$

By analogous means we obtain the equations

$$(2.38) \quad \begin{aligned} & \sum_{i=1}^m \left[\int_G \Delta z_i L_{it} y dx dt + \int_0^T \int_\Gamma \Delta z_i P_i y d\sigma dt \right] \\ & = \sum_{i=1}^m \left[\int_G \Delta \frac{\partial H}{\partial z_i} \Delta z_i dx dt + \int_0^T \int_\Gamma \Delta \frac{\partial h}{\partial z_i} \Delta z_i d\sigma dt \right], \end{aligned}$$

$$\begin{aligned}
(2.39) \quad & \sum_{i=1}^m \left[\int_C z_i L_{it} \Delta y \, dx \, dt + \int_0^T \int_{\Gamma} z_i P_i \Delta y \, d\sigma \, dt \right] \\
&= \sum_{i=1}^m \left[\int_C \left(\frac{\partial H}{\partial y_i} \Delta y_i + \sum_{k=1}^n \frac{\partial H}{\partial y_{ik}} \Delta y_{ik} \right) dx \, dt \right. \\
&\quad \left. + \int_0^T \int_{\Gamma} \frac{\partial h}{\partial y_i} \Delta y_i \, d\sigma \, dt \right] + \Delta \bar{S}_2,
\end{aligned}$$

where

$$\begin{aligned}
\Delta \bar{S}_2 = \sum_{i=1}^m \left[\int_G z_i(T, x) \Delta y_i(T, x) \, dx - \int_C \beta_i(t, x) \Delta y_i(t, x) \, dx \, dt \right. \\
\left. - \int_0^T \int_{\Gamma} \gamma_i(t, x) \Delta y_i(t, x) \, d\sigma \, dt \right].
\end{aligned}$$

From formulas (2.36)–(2.39) we get

$$\begin{aligned}
\Delta K_2[\omega] = \int_C \left[\sum_{i=1}^N \left(\frac{\partial H}{\partial w_i} + \frac{1}{2} \Delta \frac{\partial H}{\partial w_i} \right) \Delta w_i - \Delta H \right] dx \, dt \\
+ \int_0^T \int_{\Gamma} \left[\sum_{i=1}^{2m} \left(\frac{\partial h}{\partial p_i} + \frac{1}{2} \Delta \frac{\partial h}{\partial p_i} \right) \Delta p_i - \Delta h \right] d\sigma \, dt + \Delta \bar{S}_2 - \eta_2,
\end{aligned}$$

where η_2 is defined by (2.28). We apply Taylor's formula to the functions $H, h, \partial H/\partial w_i$ and $\partial h/\partial p_i$ and restrict ourselves in the expansions to second order terms. Then we shall have

$$\begin{aligned}
(2.40) \quad \Delta K_2[\omega] = \Delta \bar{S}_2 - \int_C [H(t, x, w(t, x), u_1) - H(t, x, w(t, x), u(t, x))] \, dx \, dt \\
- \int_0^T \int_{\Gamma} [h(t, x, p(t, x), v_1) - h(t, x, p(t, x), v(t, x))] \, d\sigma \, dt - \eta_8 - \eta_9,
\end{aligned}$$

where

$$\begin{aligned}
(2.41) \quad \eta_8 = \frac{1}{2} \left\{ \sum_{i=1}^N \int_C \left(\frac{\partial H(t, x, w, u_1)}{\partial w_i} - \frac{\partial H(t, x, w, u)}{\partial w_i} \right) \Delta w_i \, dx \, dt \right. \\
\left. + \sum_{i=1}^{2m} \int_0^T \int_{\Gamma} \left(\frac{\partial h(t, x, p, v_1)}{\partial p_i} - \frac{\partial h(t, x, p, v)}{\partial p_i} \right) \Delta p_i \, d\sigma \, dt \right\} \\
\eta_9 = \frac{1}{2} \left\{ \sum_{i,k=1}^N \int_C \left[\frac{\partial^2 H(t, x, \bar{w}, u_1)}{\partial w_i \partial w_k} - \frac{\partial^2 H(t, x, \tilde{w}, u_1)}{\partial w_i \partial w_k} \right] \Delta w_i \Delta w_k \, dx \, dt \right. \\
\left. + \sum_{i,k=1}^{2m} \int_0^T \int_{\Gamma} \left[\frac{\partial^2 h(t, x, \bar{p}, v_1)}{\partial p_i \partial p_k} - \frac{\partial^2 h(t, x, \tilde{p}, v_1)}{\partial p_i \partial p_k} \right] \Delta p_i \Delta p_k \, d\sigma \, dt \right\} \\
\bar{w}, \tilde{w} \in [w, w_1], \quad \bar{p}, \tilde{p} \in [p, p_1].
\end{aligned}$$

The increment of functional S_2 in (2.33) will have the form

$$\begin{aligned}
 \Delta S_2 = & \sum_{i=1}^m \left\{ \int_G \left[\alpha_i(x) \frac{\partial y_i(T, x, \omega)}{\partial T} + \beta_i(T, x) y_i(T, x, \omega) \right] dx \right. \\
 & + \int_{\Gamma} \gamma_i(T, x) y_i(T, x, \omega) d\sigma \Big\} \Delta T + \sum_{i=1}^m \left\{ \int_G \alpha_i(x) \Delta_{\omega} y_i(T, x, \omega) dx \right. \\
 (2.42) \quad & + \int_0^T \int_G \beta_i(t, x) \Delta y_i(t, x, \omega) dx dt \\
 & \left. + \int_0^T \int_{\Gamma} \gamma_i(t, x) \Delta y_i(t, x, \omega) d\sigma dt \right\} - \eta_{10},
 \end{aligned}$$

where

$$\begin{aligned}
 \eta_{10} = & \sum_{i=1}^m \left\{ \int_G \left(\alpha_i(x) \sum_{k=1}^p \left[\frac{\partial y_i(T, x, \omega)}{\partial T} - \frac{\partial y_i(\bar{t}_k, x, \omega)}{\partial T} \right] \Delta t_k \right. \right. \\
 & + [\beta_i(T, x) y_i(T, x, \omega) - \beta_i(\bar{t}, x) y_i(\bar{t}, x, \omega_1)] \Delta T \Big) dx \\
 & \left. + \Delta T \int_{\Gamma} [\gamma_i(T, x) y_i(T, x, \omega) - \gamma_i(\bar{t}, x) y_i(\bar{t}, x, \omega_1)] d\sigma \right\}.
 \end{aligned}$$

The remaining terms in functional (2.33) have the same form as the corresponding terms in functional (2.18). Therefore, their increments are determined by (2.31) in which the functions $y_i(t, x)$ and $z_i(t, x)$ must now be taken as the solutions of the second boundary value problems (2.4)–(2.6) and (2.9)–(2.11). Thus, from relations (2.34), (2.40), (2.31) and (2.42) we get

$$\begin{aligned}
 & - \int_G [H(t, x, w(t, x), u_1) - H(t, x, w(t, x), u)] dx dt \\
 (2.44) \quad & - \int_0^T \int_{\Gamma} [h(t, x, p(t, x), v_1) - h(t, x, p(t, x), v)] d\sigma dt \geq \eta,
 \end{aligned}$$

where

$$\eta = \eta_2 + \sum_{i=4}^{10} \eta_i.$$

If we utilize the notations in (2.13) and (2.14), the latter inequality can be written as

$$(2.45) \quad -\Delta J_1[u] - \Delta J_2[v] \geq \eta.$$

Thus, we obtain similar forms for the estimates of the increments of functionals J_1 and J_2 in the first and second boundary value problems

(2.4)–(2.6). By applying the same methods as used in Part I, we investigate below only the second boundary value problem (2.4)–(2.6). The first boundary value problem can be treated analogously.

2.3. Proof of Theorem 6. Let the control $\omega(t, x) = \{u(t, x), v(t, x)\}$, $0 \leq t \leq T$, $x \in G + \Gamma$, transferring the system from the state (2.5) to the state (2.8), be optimal relative to functional S_2 in the second boundary value problem (2.4)–(2.6). Let $y(t, x, \omega)$ denote the corresponding solution of this problem. We suppose that the theorem is false. Then for any system of functions $z_1(t, x), \dots, z_m(t, x)$, $a_1(x), \dots, a_j(x)$ and constants b_1, \dots, b_k satisfying (2.9) with supplementary conditions (2.10), (2.11₂) and (2.17), the control $\omega(t, x)$ does not satisfy the maximum condition relative to the functions $z(t, x)$. This implies that for every such system of functions and constants we can find an admissible control $\omega_1(t, x)$, $0 \leq t \leq T_1$, $x \in G + \Gamma$, transferring the system from the state (2.5) to the state (2.8), which satisfies one of the inequalities

$$(2.46_1) \quad \Delta J_1[u] = \int_0^\tau \int_G [H(t, x, w(t, x), u_1) - H(t, x, w(t, x), u)] dx dt > 0,$$

$$(2.46_2) \quad \Delta J_2[v] = \int_0^\tau \int_\Gamma [h(t, x, p(t, x), v_1) - h(t, x, p(t, x), v)] d\sigma dt > 0,$$

where $\tau = \min \{T_1, T_2\}$.

Let us take any one such system of functions $z_1(t, x), \dots, z_m(t, x)$, $a_1(x), \dots, a_j(x)$ and constants b_1, \dots, b_k . To be specific, we assume that for this system the controls $\omega(t, x)$ and $\omega_1(t, x)$ mentioned above satisfy inequality (2.46₂) but do not satisfy inequality (2.46₁). Let $y(t, x, \omega_1)$ denote the solution of the second boundary value problem (2.4)–(2.6). Further, let $z(t, x, \omega_1)$ denote the solution of the second boundary value problem (2.9)–(2.11), corresponding to the control ω_1 , the functions $a_1(x), \dots, a_j(x)$, and the constants b_1, \dots, b_k .

It follows from inequality (2.46₂) that in the region σ , $0 \leq t \leq \tau$, $x \in \Gamma$, there exists a point (α, β) at which

$$h(\alpha, \beta, p(\alpha, \beta), v_1(\alpha, \beta)) - h(\alpha, \beta, p(\alpha, \beta), v(\alpha, \beta)) > 0.$$

The functions $h(t, x, p(t, x), v_1(t, x))$ and $h(t, x, p(t, x), v(t, x))$ are piecewise-continuous in the region being considered. Therefore, for an arbitrarily small ϵ_1 , $\epsilon_1 > 0$, we can find a number δ and a closed region σ_{ϵ_1} (whose area is ϵ_1) containing the point (α, β) such that

$$(2.47) \quad h(t, x, p(t, x), v_1(t, x)) - h(t, x, p(t, x), v(t, x)) \geq \delta > 0$$

for all points $(t, x) \in \sigma_{\epsilon_1}$. It is obvious that this inequality is satisfied for any ϵ in the segment $0 \leq \epsilon \leq \epsilon_1$.

We take a certain ϵ , $0 < \epsilon \leq \epsilon_1$, and we construct the auxiliary control $\omega^\epsilon(t, x) = \{u^\epsilon(t, x), v^\epsilon(t, x)\}$, $0 \leq t \leq T_\epsilon$, $x \in \Gamma$, such that it satisfies the following conditions:

(i) it is defined in the region C_1 , $0 \leq t \leq \theta = \min\{T, T_\epsilon\}$, $x \in G + \Gamma$, by the formula

$$u^\epsilon(t, x) = u(t, x),$$

$$v^\epsilon(t, x) = \begin{cases} v_1 & \text{when } (t, x) \in \sigma_\epsilon, \\ v(t, x) & \text{when } (t, x) \in \sigma_1 \setminus \sigma_\epsilon, \end{cases}$$

where $\sigma_1 = \sigma$, $0 \leq t \leq \theta$, $x \in \Gamma$;

(ii) for $t > \theta$ the control ω^ϵ is defined arbitrarily except that it should be admissible and should transfer the system from the state (2.5) to the state (2.8);

(iii) the inequality $|T - T_\epsilon| < L\epsilon$ is fulfilled, where the number L is independent of ϵ .

Such a control exists by virtue of the condition that the class of admissible controls be complete. Then, applying inequality (2.47) we have

$$\Delta J_1[u] = 0,$$

$$(2.48) \quad \Delta J_2[v] = \int_0^\theta \int_\Gamma [h(t, x, p(t, x), v^\epsilon) - h(t, x, p(t, x), v)] d\sigma dt \geq \delta\epsilon > 0.$$

On the other hand, we can show that the right-hand side of inequality (2.44) is a small quantity of order $o(\epsilon)$ ($o(\epsilon)/\epsilon \rightarrow 0$ as $\epsilon \rightarrow 0$).

Indeed, the functions $\Delta y_i = y_i(t, x, \omega^\epsilon) - y_i(t, x, \omega)$ and $\Delta z_i = z_i(t, x, \omega^\epsilon) - z_i(t, x, \omega)$ form the solution of the system of equations (2.19) with supplementary conditions (2.20) and (2.35) where ω^ϵ should be substituted for ω_1 . In [19] it was shown that the functions Δy_i satisfy the inequalities

$$|\Delta g_i| \leq \int_0^\theta \int_\Gamma M(t, \tau, x, \xi) \sum_{s=1}^q |\Delta v_s| d\xi d\tau, \quad \Delta v_s = v_s^\epsilon - v_s,$$

where g is the vector with components $y_1, \dots, y_m, y_{11}, \dots, y_{mn}$, and M is a scalar function of the type of a Green's function of boundary value problems for parabolic equations. Since by construction Δv_s is nonzero only in the region σ_ϵ and since this region is bounded,

$$(2.49) \quad |\Delta g_i| = \eta_i(t, x),$$

where the η_i converge uniformly to zero with respect to t and x as $\epsilon \rightarrow 0$.

The functions $z_i(t, x)$ satisfy (2.9) with supplementary conditions (2.10) and (2.11₂). Consequently, these functions satisfy the following system of integro-differential equations [12, pp. 90–96]:

$$\begin{aligned}
 z(t, x, \omega) &= \int_G K_{11}(t, T, x, \xi) z(T, \xi) d\xi \\
 &+ \int_t^T \int_G K_{11}(t, \tau, x, \xi) A(\tau, \xi, w) d\xi d\tau \\
 &+ \int_t^T \int_\Gamma K_{12}(t, \tau, x, \xi) \phi(\tau, \xi) d\xi d\sigma d\tau, \quad x \in G, 0 \leq t \leq T, \\
 (2.50) \quad \phi(t, X, \omega) &= -a(t, X, w) + \int_G K_{22}(t, T, x, \xi) z(T, \xi) d\xi \\
 &+ \int_t^T \int_G K_{21}(t, \tau, X, \xi) A(\tau, \xi, w) d\xi d\tau \\
 &+ \int_t^T \int_\Gamma K_{22}(t, \tau, X, \xi) \phi(\tau, \xi) d\xi d\sigma d\tau, \quad x \in \Gamma, 0 \leq t \leq T,
 \end{aligned}$$

where the matrices K_{ik} are of the same type as the Green's matrix of the boundary value problem being considered, $z(T, x)$ is the initial value of the desired vector function and is determined by (2.10), while the vectors $A = (A_1, \dots, A_r)$ and $a = (a_1, \dots, a_m)$ are given by

$$\begin{aligned}
 A_i &= -\frac{\partial H}{\partial y_i} + \sum_{k=1}^n \frac{\partial H}{\partial y_{ik}} + \beta_i(t, x), \\
 a_i &= \frac{\partial h}{\partial y_i} + \sum_{k=1}^n \frac{\partial H}{\partial y_{ik}} X_k - \gamma_i(t, x), \quad i = 1, \dots, m.
 \end{aligned}$$

The functions $z(t, x, \omega^\epsilon)$ and $\phi(t, X, \omega^\epsilon)$ satisfy analogous equations where the initial conditions should be taken at $t = T_\epsilon$. For definiteness we shall take $T = T_\epsilon$ and, consequently, $\theta = T$. We denote the corresponding matrices by K_{ik}^ϵ .

Since the functions $\partial H / \partial y_i$, $\partial H / \partial y_{ik}$ and $\partial h / \partial y_i$ satisfy a Lipschitz condition, with due regard to (2.49) we get

$$\begin{aligned}
 |\Delta A_i| &\leq R \sum_{k=1}^m |\Delta z_k| + \eta_{1i}(t, x), \quad x \in G, \\
 (2.51) \quad |\Delta a_i| &\leq R \sum_{k=1}^m |\Delta z_k| + \eta_{2i}(t, x), \quad x \in \Gamma, 0 \leq t \leq T,
 \end{aligned}$$

where R is a constant and the $\eta_{ki}(t, x)$ converge uniformly to zero with respect to t and x as $\epsilon \rightarrow 0$.

By introducing the function $\phi(t, x, \omega)$ defined by the second equation in (2.50) into the right-hand side of the same equation and by repeating this operation several times, we obtain

$$\begin{aligned}
 \phi(t, X, \omega) = & -a(t, X, w) + \int_t^T \int_{\Gamma} \sum_{i=1}^{n-1} K^i(t, \tau, X, \xi) a(\tau, \xi, w) d\xi \sigma d\tau \\
 (2.52) \quad & + \int_G K_2^n(t, T, X, \xi) z(T, \xi) d\xi + \int_t^T \int_G K_n(t, \tau, X, \xi) A(\tau, \xi, w) d\xi d\tau \\
 & + \int_t^T \int_{\Gamma} K^n(t, \tau, X, \xi) \phi(\tau, \xi, \omega) d\xi \sigma d\tau,
 \end{aligned}$$

where

$$\begin{aligned}
 K_n(t, \tau, X, \xi) = & K_{n-1}(t, \tau, X, \xi) \\
 & + \int_t^{\tau} \int_{\Gamma} K_{n-1}(t, \alpha, X, \beta) K^0(\alpha, \tau, \beta, \xi) d\beta \sigma d\alpha, \\
 K^n(t, \tau, X, \xi) = & \int_t^{\tau} \int_{\Gamma} K^{n-1}(t, \alpha, X, \beta) K_0(\alpha, \tau, \beta, \epsilon) d\beta \sigma d\alpha, \\
 K_0 = & K_{21}, \quad K^0 = K_{22}, \\
 K_2^n(t, T, X, \eta) = & K_{22}(t, T, X, \eta) \\
 & + \int_t^{\tau} \int_{\Gamma} K_{22}(t, T, \xi, \eta) \sum_{i=1}^{n-1} K^i(t, \tau, X, \xi) d\xi \sigma d\tau, \quad i = 1, 2, \dots.
 \end{aligned}$$

An analogous equation is obtained for the function $\phi(t, X, \omega^\epsilon)$. In these equations the number n is chosen so large that the corresponding kernels K^n are bounded. This can be done by virtue of the well-known estimates of Green's matrix and of its derivatives [12, p. 92]. Therefore, taking (2.49) and (2.51) into account we obtain the inequality

$$\begin{aligned}
 W(t) \leq & P \int_t^T W(\tau) d\tau + \int_t^T \int_G Q_n(t, \tau, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi d\tau \\
 (2.53) \quad & + \int_t^T \int_{\Gamma} R_n(t, \tau, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi \sigma d\tau \\
 & + P_1 \int_{\Gamma} \sum_{i=1}^m |\Delta z_i(t, X)| d\sigma + \delta(t),
 \end{aligned}$$

where the functions Q_n and R_n have weak singularities, $\delta(t)$ converges uniformly to zero with respect to t as $\epsilon \rightarrow 0$, and

$$(2.54) \quad W(t) = \int_{\Gamma} \sum_{i=1}^m |\phi_i(t, X, \omega^\epsilon) - \phi_i(t, X, \omega)| d\sigma.$$

We introduce the notation:

$$\begin{aligned}
W_k(t) &= \int_t^T W_{k-1}(\tau) d\tau, & Q_{nk} &= \int_t^\tau Q_{n,k-1}(t, \tau, \xi) dt, \\
W_0(t) &= W(t), & Q_{n0} &= Q_n, \\
(2.55) \quad R_{nk} &= \int_t^\tau R_{n,k-1}(t, \tau, \xi) dt, \\
R_{n1} &= \int_t^\tau R_1(t, \tau, \xi) dt + P_1, \\
\delta^k(t) &= \int_t^T \frac{(\tau - t)^{k-1}}{k-1} \delta_i(\tau) d\tau.
\end{aligned}$$

Integrating inequality (2.53) k times we get

$$\begin{aligned}
(2.56) \quad W_k(t) &\leq P \int_t^T W_k(\tau) d\tau + \int_t^T \int_G Q_{nk}(t, \tau, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi d\tau \\
&\quad + \int_t^T \int_\Gamma R_{nk}(t, \tau, \xi) \sum_{i=1}^m |\Delta z_i| d\sigma d\tau + \delta^k(t).
\end{aligned}$$

We choose the number k so large that the functions Q_{nk} and R_{nk} are bounded for $0 \leq t \leq \tau \leq T$, $\xi \in G + \Gamma$, and we set

$$\begin{aligned}
Q(t) &= \int_t^T \left[\int_G \sum_{i=1}^m |\Delta z_i| \max_{t \leq \theta \leq T} Q_{nk}(\theta, \tau, \xi) d\xi \right. \\
&\quad \left. + \int_\Gamma \sum_{i=1}^m |\Delta z_i| \max_{t \leq \theta \leq T} R_{nk}(t, \tau, \xi) d\sigma \right] d\tau + \delta^{(k)}(t).
\end{aligned}$$

Then, from inequality (2.56) we get $W_k(\theta) \leq \int_\theta^T W_k(\theta) d\theta + Q(t)$.

Applying Bellman's lemma [21, p. 35] we find that $W_k(\theta) \leq P_1 Q(t)$ for $t \leq \theta \leq T$ and, consequently, $W_k(t) \leq P_1 Q(t)$, where P_1 is a specified positive constant. Therefore, from (2.53)–(2.56) we obtain

$$\begin{aligned}
(2.57) \quad W(t) &\leq \int_t^T \int_G M_1(t, \tau, \xi) \sum |\Delta z_i| d\xi d\tau \\
&\quad + \int_0^T \int_\Gamma N_1(t, \tau, \xi) \sum |\Delta z_i| d\sigma d\tau + \eta(t),
\end{aligned}$$

where M_1 and N_1 are functions of the same type as Q_n and R_n , and $\eta(t)$ converges uniformly to zero with respect to t as $\epsilon \rightarrow 0$.

If we set up (2.52) for the function $\phi(t, X, \omega^\epsilon)$ and also take conditions (2.51), (2.57) and notation (2.54) into account, we obtain

$$\begin{aligned}
 \sum_{i=1}^m |\Delta \phi_i(t, X)| &\leq P \sum_{i=1}^m |\Delta z_i(t, X)| \\
 (2.57_1) \quad &+ \int_G M_2(t, T, X, \xi) \sum_{i=1}^m |\Delta z_i(T, \xi)| d\xi \\
 &+ \int_t^T \int_G M_3(t, \tau, X, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi d\tau \\
 &+ \int_t^T \int_\Gamma N_2(t, \tau, X, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi d\tau + \eta_1(t, X),
 \end{aligned}$$

where P is a constant, M_i and N_i are functions of the same type as K_n and K^n .

The functions $z_i(T, x, \omega)$ and $z_i(T, x, \omega^\epsilon)$ are determined from conditions (2.10) and, consequently, according to condition (2.49),

$$(2.57_2) \quad \sum_{i=1}^m |\Delta z_i(T, x, \omega)| = \eta_2(x),$$

where η_2 converges uniformly to zero with respect to x as $\epsilon \rightarrow 0$.

Setting (2.50) up for the function $z(t, x, \omega^\epsilon)$ and taking (2.57₁) and (2.57₂) into account, we get

$$\begin{aligned}
 \sum_{i=1}^m |\Delta z_i(t, x)| &\leq \int_t^T \int_G M_4(t, \tau, x, \xi) \sum_{i=1}^m |\Delta z_i(\tau, \xi)| d\xi d\tau \\
 &+ \int_t^T \int_\Gamma N_3(t, \tau, x, \xi) \sum_{i=1}^m |\Delta z_i| d\xi d\tau + \eta_3.
 \end{aligned}$$

Hence by the very same method used to obtain inequality (2.57₂) we find that

$$|\Delta z_i(t, x)| = \eta_i(t, x), \quad 0 \leq t \leq T, x \in G + \Gamma.$$

The functions $\Delta z_{ik} = \partial \Delta z_i / \partial x_k$ satisfy analogous conditions.

Just as in Part I it can be proved that the quantity η occurring in the right-hand side of inequality (2.44) is of a higher order of smallness than ϵ , i.e.,

$$\eta = o(\epsilon), \quad \frac{o(\epsilon)}{\epsilon} \rightarrow 0 \quad \text{as } \epsilon \rightarrow 0.$$

This relation (together with inequalities (2.46)) contradicts the fact that the control $\omega(t, x)$ satisfies inequality (2.44).

2.4. Problems with other optimality criteria. Theorems 5 and 6 give us the necessary optimality conditions for problems where we are required to minimize linear functionals S_1 or S_2 . However, the method we have pro-

posed, together with the results of Part I, allow us to consider more general cases when nonlinear functionals or time-optimality serve as the optimality criterion. To simplify the subsequent formulations we shall treat the case when the system of equations (2.4) has the form

$$(2.58) \quad \frac{\partial y_i}{\partial t} - k_i^2 \frac{\partial^2 y_i}{\partial x^2} = f_i(t, x, y, y_x, u),$$

$$0 \leq t \leq T, \quad 0 \leq x \leq X, \quad i = 1, \dots, m,$$

where the k_i^2 are constants and the f_i satisfy the same conditions as before. We take the initial conditions in the form

$$(2.59) \quad y_i(0, x) = a_i(x),$$

where the $a_i(x)$ are continuous functions. We take the boundary conditions in one of the following forms

$$(2.60_1) \quad y_i(t, 0) = \phi_i^0(t), \quad y_i(t, X) = \phi_i^1(t), \quad i = 1, \dots, m,$$

or

$$(2.60_2) \quad \frac{\partial y_i(t, 0)}{\partial x} = \psi_i^0(t, y, v^0), \quad \frac{\partial y_i(t, X)}{\partial x} = \phi_i(t, y, v^1), \quad i = 1, \dots, m,$$

where the $\phi_i^k(t)$ are continuous functions of the variable t , while the $\psi_i^k(t, y, v^k)$ are continuous in t and twice differentiable in y and v^k .

As admissible controls we take the functions $u(t, x)$ and $\omega(t, x) = \{u(t, x), v^0(t), v^1(t)\}$ with the same properties as in the problem considered above, while as the optimality criterion we have

$$S = \int_0^T \int_0^X f_0(t, x, y, y_x, u) \, dx \, dt,$$

in which the quantity T can depend on the controls. The terminal state of the system is given by the relations

$$\Phi_\alpha(T, x, y(T, x)) = 0, \quad \alpha = 1, \dots, j,$$

$$\int_0^X \Psi_\beta(T, x, y(T, x)) \, dx = c_\beta, \quad \beta = 1, \dots, k,$$

where $j + k \leq m$, Φ_α and Ψ_β satisfy the same conditions as the functions corresponding to them in the problems analyzed above.

In order to apply the method we have presented for solving the optimal problem, we introduce the auxiliary variable y_0 by setting

$$(2.62) \quad \frac{\partial^2 y_0}{\partial t \partial x} = f_0(t, x, y, y_x, u), \quad y_0(t, 0) = y_0(0, x) = 0.$$

Then, to every admissible control in the first (second) boundary value problem (2.58)–(2.60), which transfers the system from the state (2.59) to the state (2.61), there corresponds a unique function $y_0(t, x)$ defined by (2.62), while the functional S takes the form $S = y_0(T, X)$ and is defined on the solutions y_0, \dots, y_m .

We should make use of the results of Part I to solve the optimal control problem thus obtained, since one of the equations of the system being studied (namely (2.62)) is hyperbolic and the variable y_0 does not enter into (2.58). For definiteness we shall consider only the second boundary value problem (2.58)–(2.60).

The corresponding adjoint boundary value problem will be

$$(2.63) \quad \frac{\partial^2 z_0}{\partial t \partial x} = 0, \quad \frac{\partial z_0(t, X)}{\partial t} = \frac{\partial z_0(T, x)}{\partial x} = 0, \quad z_0(T, X) = -1,$$

$$\frac{\partial z_i}{\partial t} + k_i^2 \frac{\partial^2 z_i}{\partial x^2} = -\frac{\partial H}{\partial y_i} + \frac{d}{dx} \left(\frac{\partial H}{\partial y_{ix}} \right),$$

$$(2.64) \quad a_i^2 \frac{\partial z_i^0}{\partial x} = \frac{\partial h^0}{\partial y_i^0} + \frac{\partial H^0}{\partial y_{ix}}, \quad a_i^2 \frac{\partial z_i^1}{\partial x} = \frac{\partial h^1}{\partial y_i^1} + \frac{\partial H^1}{\partial y_{ix}},$$

$$(2.65) \quad z_i(T, x) = -\sum_{\alpha=1}^j a_\alpha(x) \frac{\partial \Phi_\alpha}{\partial y_i} - \sum_{\beta=1}^k b_\beta \left(\frac{\partial \Psi_\beta}{\partial y_i} - \frac{d}{dx} \left[\frac{\partial \Psi_\beta}{\partial y_{ix}} \right] \right),$$

$$(2.66) \quad \int_0^X \left[\sum_{\alpha=1}^j a_\alpha(x) \frac{d\Phi_\alpha}{dT} + \sum_{\beta=1}^k b_\beta \frac{d\Psi_\beta}{dT} - f_0(T, x, y(T, x), y_x(T, x), u(T, x)) \right] dx = 0,$$

where

$$H = \sum_{i=0}^m z_i f_i(t, x, y, y_x, u),$$

$$h^0 = \sum_{i=1}^m k_i^2 z_i^0 \psi_i^0(t, y, v^0),$$

$$h^1 = \sum_{i=1}^m k_i^2 z_i^1 \psi_i^1(t, y, v^1),$$

$$z_i^0 = z_i(t, 0), \quad z_i^1 = z_i(t, 1), \quad H^0 = H|_{x=0}, \quad H^1 = H|_{x=1}.$$

Hence we find that $z_0(t, x) = -1$ and, consequently,

$$H = \sum_{i=1}^m z_i f_i(t, x, y, y_x, u) - f_0.$$

Thus, we have obtained the following theorem.

THEOREM 7. Let the admissible control $\omega(t, x) = \{u(t, x), v^0(t), v^1(t)\}$,

$0 \leq t \leq T$, $x \in [0, X]$, transfer the system from the state (2.59) to the state (2.61), and let $y(t, x)$ be the corresponding solution of the second boundary value problem (2.58)–(2.60). In order that $y(t, x)$ and $\omega(t, x)$ be optimal relative to functional S , it is necessary that there exist functions $z(t, x)$, $a(x)$ and constants b_1, \dots, b_k such that:

(i) the functions $y(t, x)$, $z(t, x)$, $\omega(t, x)$, $a(x)$, and the constants b_1, \dots, b_k form the solution of (2.58) and (2.63) with supplementary conditions (2.59), (2.60₂), (2.64)–(2.66);

(ii) the control $\omega(t, x)$ satisfies a maximum condition relative to the function $z(t, x)$.

In order to obtain the optimality conditions in the time-optimal problem we must take $f_0 \equiv 1$ in Theorem 7.

Analogously we can find the optimality conditions for problems where one of the following expressions,

$$f_0(T, x, y(T, x)), \quad \int_0^T f_0(t, X, y(t, X)) dx, \quad \int_0^X f_0(T, x, y(T, x)) dx,$$

is chosen as the functional to be minimized.

The same procedure is applicable for the solution of problems with an arbitrary finite number of independent variables x_1, \dots, x_n for which the optimality criterion appears as a multidimensional integral. In this case we must use the method for investigating control processes in systems whose behavior is described by a multidimensional Goursat problem [19].

REFERENCES

- [1] I. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [2] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.
- [3] A. G. BUTKOVSKII AND A. YA. LERNER, *Optimal control of distributed parameter systems*, Automat. Remote Control, 21 (1960), pp. 472–477.
- [4] K. A. LUR'E, *The Mayer-Bolza problem for multiple integrals and the optimization of the performance of distributed-parameter systems*, J. Appl. Math. Mech., 27 (1963), pp. 1284–1299.
- [5] T. K. SIRAZETDINOV, *On the theory of optimal processes with distributed-parameters*, Automat. Remote Control, 25 (1964), pp. 431–440.
- [6] YU. M. VOLIN AND G. M. OSTROVSKII, *On an optimal problem*, Ibid., 25 (1964), pp. 1271–1277.
- [7] YU. V. EGOROV, *Certain problems in optimal control theory*, USSR Comput. Math. and Math. Phys., 3 (1963), pp. 1209–1232.
- [8] ———, *On hyperbolic equations with discontinuous coefficients*, Soviet Math. Dokl., 1 (1960), pp. 1095–1098.
- [9] A. I. EGOROV, *Optimal control of processes in certain distributed-parameter systems*, Automat. Remote Control, 25 (1964), pp. 557–566.

- [10] G. A. BLISS, *Lectures on the calculus of variations*, University of Chicago Press, Chicago, 1946.
- [11] C. MIRANDA, *Equazioni alle Derivate Parziali di Tipo Ellittico*, Springer-Verlag, Berlin, 1955.
- [12] T. YA. ZAGORSKII, *Mixed Problems for Systems of Parabolic Partial Differential Equations*, Izd-vo L'vovsk. Un-ta, 1961.
- [13] O. A. OLEINIK, *Boundary value problems for linear elliptic and parabolic equations with discontinuous coefficients*, Amer. Math. Soc. Transl., Ser. 2, 42 (1964), pp. 175–194.
- [14] L. I. KAMYNIN, *On the solutions of boundary value problems for a parabolic equation with discontinuous coefficients*, Soviet Math. Dokl., 2 (1961), pp. 1043–1046.
- [15] ———, *The method of thermal potentials for parabolic equations with discontinuous coefficients*, Sibirsk. Mat. Zh., 4 (1963), pp. 1071–1105.
- [16] YU. V. EGOROV, *Some optimal control problems*, Soviet Math. Dokl., 3 (1962), pp. 1080–1084.
- [17] A. G. BUTKOVSKII, *Optimal control theory in distributed-parameter systems*, Doctoral dissertation, Institute of Automation and Remote Control, Acad. Sci. USSR, 1963.
- [18] A. M. IL'IN, A. S. KALASHNIKOV AND O. A. OLEINIK, *Linear second-order parabolic equations*, Russian Math. Surveys, 17 (1962), pp. 1–143.
- [19] A. I. EGOROV, *Optimal processes in distributed-parameter systems and certain problems of invariance theory*, Izv. Akad. Nauk SSSR Ser. Mat., 29 (1965), pp. 1205–1260.
- [20] V. V. NEMYTSKII AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, 1960.
- [21] R. BELLMAN, *Stability Theory of Differential Equations*, McGraw-Hill, New York, 1953.

ON THE CLOSURE AND CONVEXITY OF ATTAINABLE SETS IN FINITE AND INFINITE DIMENSIONS*

H. HERMES†

Introduction. Much of the mathematical theory of optimal control deals with a system of differential equations

$$(1) \quad \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x^0,$$

where the function $u: [0, T] \rightarrow E^n$ (E^n denoting Euclidean n -dimensional space), termed the control, may be chosen from a control set $\Omega \subset \mathcal{L}_\infty[0, T]$. (If we say that a vector-valued function is in $\mathcal{L}_p[0, T]$, we shall mean that each of its components is in this space.) If we assume conditions on f , such that for a choice $u \in \Omega$ a unique solution $x(\cdot, u)$ of (1) exists, the set of derivatives of solutions, i.e., the set

$$(2) \quad \mathcal{F} = \{f(\cdot, x(\cdot, u), u(\cdot)): u \in \Omega\},$$

may usually be considered in some Lebesgue space. The attainable set $\mathcal{A}(t)$, for some $t \in [0, T]$, is defined as the set of points attainable at time t by solutions of (1) for all possible choices of control $u \in \Omega$, i.e.,

$$(3) \quad \mathcal{A}(t) = \{x^0 + \int_0^t z(\tau) d\tau: z \in \mathcal{F}\}.$$

The existence theorems which depend on the compactness of $\mathcal{A}(t)$ may then be viewed as follows: Under what conditions on \mathcal{F} will its image in E^n under the linear operator L defined by $Lz = \int_0^t z(\tau) d\tau$, be compact? We shall show, for instance, that the Filippov existence theorem [1] can be interpreted as stating that, with his assumptions, \mathcal{F} is a weak* compact subset of $\mathcal{L}_\infty[0, t]$, hence its image under the weak* continuous map L is compact.

While the most natural situation to obtain $\mathcal{A}(t)$ compact would be to seek a topological space in which \mathcal{F} is compact and L continuous, the fact that, in general, \mathcal{F} will be a subset of an infinite-dimensional space while L has finite-dimensional range suggests the possibility that, even if \mathcal{F} is not compact in its chosen space, the range of \mathcal{F} under L may be compact. In particular, this situation is illustrated by the theorem of Lyapunov on the range of a vector measure. Specifically, let \mathcal{F} be that subset of $\mathcal{L}_\infty[0, T]$,

* Received by the editors October 26, 1966.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80302. This research was supported in part by the National Aeronautics and Space Administration under Contract NAS8-11264.

such that $z \in \mathcal{F}$ implies $z_i(t)$ is either zero or one for all $t \in [0, T]$, $i = 1, 2, \dots, n$. Since \mathcal{F} is not convex, it certainly is not weak* compact nor is it compact in the norm topology of $\mathcal{L}_\infty[0, T]$, yet the Lyapunov theorem yields the fact that its image under L is compact and convex in E^n . This has been used to obtain existence theorems for linear systems (see [2], [3]).

There have been many recent generalizations of Lyapunov's theorem (e.g., [4]); it is natural to ask whether it could be applied to a nonlinear system in such a way to yield \mathcal{F} not weak* compact in $\mathcal{L}_\infty[0, T]$ but still so that its image under L is compact (and convex). This question will be pursued.

It should be noted that the use of the Lyapunov theorem in [2], [3] depended on the fact that the control set Ω was defined, as a subset of $\mathcal{L}_\infty[0, T]$, by giving the values which an element $u \in \Omega$ could assume at each $t \in [0, T]$. One might also consider Ω given merely as a subset of $\mathcal{L}_\infty[0, T]$. We shall show, using a construction of Klee [5], that it is possible for Ω to be a closed, bounded, convex, subset of $\mathcal{L}_\infty[0, T]$, such that the attainable set $\mathcal{A}(t)$ for a linear system

$$(4) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = x^0,$$

with the components of A, B in $\mathcal{L}_1[0, T]$, is *not* closed.

We shall also discuss possible infinite-dimensional analogues of the Lyapunov theorem.

1. Statement of results. The control set Ω will be assumed to be given in either of the two following ways:

(i) For each $t \in [0, T]$ let $U(t)$ be any subset of E^r and $\Omega = \{u \in \mathcal{L}_\infty[0, T] : u(t) \in U(t)\}$. We assume the sets $U(t)$ are contained in some fixed bounded sphere S in E^r for $t \in [0, T]$.

(ii) Ω is a bounded subset of r vector-valued functions with components in $\mathcal{L}_\infty[0, T]$. (Again we assume the values of elements of Ω are contained in S .)

There will be a need to consider \mathcal{L}_∞ with its norm topology, its weak topology, and its weak* topology (or \mathcal{L}_1 topology of \mathcal{L}_∞). For writing ease both notations for the weak* topology will be used.

The assumptions on f in (1) will be the following:

1. f is continuous on $[0, T] \times E^n \times S$ and once continuously differentiable in the x argument, unless explicitly stated otherwise.

2. There exists a constant $c > 0$ such that $x \cdot f(t, x, u) \leq c[1 + |x|^2]$ for all t, x, u in the domain of definition of f . (This prevents finite escape time.)

With these assumptions, for each $u \in \Omega$, (1) has a unique solution defined on $[0, T]$ which will be denoted $x(\cdot, u)$.

In addition to the sets \mathfrak{F} and $\mathfrak{A}(t)$ defined in the Introduction, we define, for the case Ω having the representation (i), the sets

$$F(t) = \{f(t, a, \sigma) : a \in \mathfrak{A}(t), \sigma \in U(t)\},$$

and

$$\mathfrak{B}(t) = \{x^0 + \int_0^t z(\tau) d\tau : z \text{ measurable, } z(\tau) \in F(\tau) \text{ for } 0 \leq \tau \leq t\}.$$

Remark 1. The set $F(t)$ is related to the "local direction cone" $\{f(t, x, \sigma) : \sigma \in U(t)\}$ and will always contain this set since it is the union of such sets taken over all x in the attainable set $\mathfrak{A}(t)$. It easily follows that we will always have $\mathfrak{A}(t) \subset \mathfrak{B}(t)$; one of the things we shall be interested in is when are these sets equal.

Remark 2. From the assumptions on f and Ω it follows that \mathfrak{F} is a bounded subset of $\mathfrak{L}_\infty[0, T]$.

We shall next summarize results. In doing so several theorems from other references will be stated; at times the statements of these may be somewhat different from the form in which they originally appeared. In these cases, the verification of the equivalence will be included in §2, where proofs of the results are given.

I [4, Theorem 1]. $\mathfrak{B}(t)$ is convex for each $t \in [0, T]$.

II [4, Theorem 4]. If $F(\tau)$ is closed for each $\tau \in [0, t]$ (our assumptions imply it is bounded), then $\mathfrak{B}(t)$ is convex and compact for each $t \in [0, T]$.

III [1, Theorem 1 and Lemma]. Suppose Ω has the representation (i) with $U(t)$ a nonempty compact subset of E^r for each $t \in [0, T]$ which is continuous in the Hausdorff topology as a function of t . Suppose further that for each t, x , $\{f(t, x, \sigma) : \sigma \in U(t)\}$ is convex. Then $F(t)$ is closed for each $t \in [0, T]$.

IV (Restatement of [1, Theorem 1]). Assume the hypotheses of III. Then \mathfrak{F} is a weak* compact subset of $\mathfrak{L}_\infty[0, T]$.

Remark. From this it immediately follows that $\mathfrak{A}(t)$ is compact. Indeed the mapping $L : \mathfrak{L}_\infty \rightarrow E^n$ defined by $Lz = \int_0^t z(\tau) d\tau$ is weak* continuous, and hence the image of \mathfrak{F} is compact.

V. Assume the hypotheses of III and that for each $\tau \in [0, T]$ and $a, a' \in \mathfrak{A}(\tau)$,

$$(7) \quad \{f(\tau, a, \sigma) : \sigma \in U(\tau)\} = \{f(\tau, a', \sigma) : \sigma \in U(\tau)\}.$$

Then $\mathfrak{A}(t) = \mathfrak{B}(t)$ for each $t \in [0, T]$.

VI (Combining II, III and V). If the hypotheses of V are satisfied, $\mathfrak{A}(t)$ is compact and convex for each $t \in [0, T]$.

Remark. Compactness of $\mathcal{A}(t)$ is essential in existence theorems, and the convexity of $\mathcal{A}(t)$ plays a role in ruling out “conjugate points” [6], thus simplifying sufficiency conditions.

The next few results pertain to the case where Ω has the representation (ii).

Let X^* be the dual of a Banach space X ; then every closed and bounded (in norm) convex set in X^* is closed in the X^{**} (or weak) topology of X^* . Also, a subset of X^* is compact in the X topology of X^* if and only if it is bounded in the norm topology and closed in the X topology (see [7, pp. 422–424]).

In [5, p. 881] Klee shows that every nonreflexive separable Banach space contains two disjoint, closed, bounded, convex sets which cannot be separated. As remarked in [5], the separability is not essential since every nonreflexive space has a nonreflexive closed separable subspace within which one could apply the result. Using Klee’s result one easily obtains the following.

VII. If X^* is a nonreflexive Banach space which is the dual of a Banach space X , it contains a closed, bounded, convex subset Ω_1 which is not closed in the X topology of X^* .

For any $y = (y_1, \dots, y_n)$ with components in \mathcal{L}_1 , let $L(y)$ denote the linear operator from \mathcal{L}_∞ to E^n defined by

$$(8) \quad L(y)u = \int_0^T y(\tau)u(\tau) d\tau$$

where we assume u is scalar valued. It will be useful, in §2, to use the functional notation $y(u)$ for $L(y)u$; this will sometimes be done.

VIII. There exists a $y \in \mathcal{L}_1[0, T]$, such that the image of the closed, bounded, convex set $\Omega_1 \subset \mathcal{L}_\infty$ under the continuous linear map $L(y)$ is not closed in E^n .

Equivalently, it readily follows from this that there exists a linear control system of the form (4) with control set Ω_1 , a closed, bounded, convex subset of $\mathcal{L}_\infty[0, T]$, for which $\mathcal{A}(T)$ is not compact, i.e., there need not be existence of an optimal control for a time-optimal problem.

The preceding deals with finite-dimensional systems; results concerning convexity of the attainable set $\mathcal{A}(t)$ are derived from I , which is an immediate consequence of the Lyapunov theorem on the range of nonatomic measures. Since the proofs of this theorem proceed by induction on the dimension, one may wonder whether it has an analogue for measures with values in a Banach space. Specifically, consider the special case of the Lyapunov theorem which states: If f is an n vector-valued function with components in $\mathcal{L}_1[0, T]$ (the measure being Lebesgue measure) while Σ denotes the σ -algebra of Lebesgue measurable subsets of $[0, T]$ with χ_E the

characteristic function of $E \in \Sigma$, then $\left\{ \int_0^T f(\tau) \chi_E d\tau : E \in \Sigma \right\}$ is closed and convex. The infinite-dimensional analogue would consider the case where f takes values in an infinite-dimensional Banach space. We shall consider the special case when the Banach space is the sequence space l_∞ , where $x \in l_\infty$ implies $x = (\xi_1, \dots)$ with $\|x\| = \sup |\xi_n|$.

IX. There exists a map $f: [0, T] \rightarrow l_\infty$ with components $f_i \in \mathcal{L}_1[0, T]$ such that $\left\{ \int_0^T f(\tau) \chi_E d\tau : E \in \Sigma \right\}$ is not convex.

Remark 1. One can easily replace l_∞ with the separable Hilbert space l_2 in IX. This result is fairly well known (for previous examples and discussions, see [8], [9]).

Remark 2. One could consider, instead of this infinite-dimensional analogue of the Lyapunov theorem, the infinite-dimensional version of I. Specifically, let C map $[0, T]$ into the set of subsets of l_∞ . Under what additional conditions on C (or the values $C(t)$) is $\left\{ \int_0^T c(\tau) d\tau : c \text{ has components in } \mathcal{L}_1[0, T], c(t) \in C(t) \right\}$ convex? Suppose there exists a subsequence $\{n_j\}$ of the sequence of positive integers with the property that $c(t) \in C(t)$ for $0 \leq t \leq T$ implies that the possible values of the components $c_{n_j}(t), c_{n_j+1}(t), \dots, c_{n_{j+1}-1}(t), j = 1, 2, \dots$, depend only on each other. Then for each j , one can apply I to the set of values $c_{n_j}, \dots, c_{n_{j+1}-1}$, may assume, and thereby obtain convexity. As an example, if $C(t)$ is the set of "vertices of the cube" in l_∞ , specifically $y \in C(t)$ implies for each i, y_i is either 0 or 1, then $\left\{ \int_0^T c(\tau) d\tau : c \text{ has components in } \mathcal{L}_1[0, T], c(\tau) \in C(\tau) \right\}$ is convex.

2. Verification of stated results.

III. $F(t)$ is closed. For Ω as given in III, Filippov's theorem [1] shows $\mathcal{A}(t)$ is compact, $U(t)$ is given compact and f is continuous, therefore $F(t)$ is compact.

IV. \mathfrak{F} is weak* closed. Let

$$z^n(\cdot) = f(\cdot, x(\cdot, u^n), u^n(\cdot)) \in \mathfrak{F}$$

and z^n converge to z in the weak* topology. We will show $z \in \mathfrak{F}$.

Since \mathfrak{F} is bounded in norm it is easily shown that z^n converges to z in the weak* topology if and only if

$$\int_0^t z^n(\tau) d\tau \rightarrow \int_0^t z(\tau) d\tau$$

for each $t \in [0, T]$ (see [7, Exercise 27, p. 342]). Letting $x(t)$

$= x^0 + \int_0^t z(\tau) d\tau$, $t \in [0, T]$, the hypotheses imply $x(\cdot, u^n)$ converges to x uniformly. But then the Filippov argument [1, proof of Theorem 1] implies x is an admissible trajectory, i.e., there exists an admissible control u such that $z(t) = \dot{x}(t) = f(t, x(t), u(t))$ for almost all $t \in [0, T]$, showing $z \in \mathfrak{F}$.

V (*Proof*). We already know $\mathfrak{A}(t) \subset \mathfrak{B}(t)$; now let $x^0 + \int_0^t z(\tau) d\tau$ be any element in $\mathfrak{B}(t)$, i.e., for each $s \in [0, t]$, $z(s) \in F(s)$. We must show $x^0 + \int_0^\tau z(s) ds$ is an admissible trajectory for $\tau \in [0, t]$.

By the representation of Ω and property (7) of the hypotheses of V, $F(s) = \{f(s, a, \sigma) : \sigma \in U(s)\}$ for any $a \in \mathfrak{A}(s)$. Pick any admissible control u^0 and let $x(\cdot, u^0)$ be its corresponding trajectory. Then for each $s \in [0, t]$, $z(s) \in \{f(s, x(s, u^0), \sigma) : \sigma \in U(s)\}$. Hence by the Filippov lemma [1], there exists an admissible control u^1 such that $z(s) = f(s, x(s, u^0), u^1(s))$ almost everywhere. Using u^1 in the place of u^0 we may proceed inductively to generate a sequence of trajectories $\{x(\cdot, u^n)\}$ and corresponding sequence of controls $\{u^{n+1}\}$, such that $z(s) = f(s, x(s, u^n), u^{n+1}(s))$ almost everywhere in $[0, t]$. From the sequence $\{x(\cdot, u^n)\}$ choose a uniformly convergent subsequence (the original sequence is easily seen to be an equicontinuous family) and for notational ease assume it is the original sequence. Define $z^n(s) = f(s, x(s, u^n), u^n(s))$. Then

$$\begin{aligned} |z^n(s) - z(s)| &= |f(s, x(s, u^n), u^n(s)) - f(s, x(s, u^{n-1}), u^n(s))| \\ &\leq K \sup \{ |x(s, u^n) - x(s, u^{n-1})|, 0 \leq s \leq t \} \end{aligned}$$

for almost all $s \in [0, t]$. But $x(\tau, u^n) = x^0 + \int_0^\tau z^n(s) ds$ is an admissible trajectory and hence, from the preceding estimate, $x(\tau, u^n)$ converges uniformly to $x^0 + \int_0^\tau z(s) ds$. Therefore, $x^0 + \int_0^\tau z(s) ds$ is the uniform limit of admissible trajectories, and by Filippov's theorem [1] it is an admissible trajectory, which completes the argument.

VII. In [5] Klee shows that every nonreflective separable Banach space contains two disjoint, closed, bounded, convex sets which cannot be separated. The separability of the space is inconsequential since, as commented in [5], every nonreflexive Banach space X^* has a separable nonreflexive closed subspace. Let E^* denote this subspace; consider A, B closed, bounded and convex in E^* and such that they cannot be separated. Then as subsets of X^* they are also closed, bounded, convex and cannot be separated by a hyperplane since $E^* \subset X^*$ implies $X^{**} \subset E^{**}$, i.e., any continuous linear functional on X^* is a continuous linear functional on E^* .

Now suppose either A or B is closed in the X topology of X^* . Then by

[7, Corollary 3, p. 424], it is compact in the X topology of X^* . This implies we have two closed, disjoint, convex sets in a locally convex linear topological space (X^* with its X topology), one of which is compact. By [7, Corollary 11, p. 418], there exists a nonzero continuous linear functional f which separates them. But if f is continuous in the weak* topology, it is continuous in the norm topology of X^* , i.e., $f \in X^{**}$. This implies f separates A and B in X^* , a contradiction. Thus neither A nor B can be closed in the X topology of X^* .

VIII (*Proof*). It suffices to consider y , as in (8), to be scalar-valued, i.e., $L(y) : \mathcal{L}_\infty \rightarrow E^1$. We will consider only real linear spaces.

Let X be a Banach space and K^* a bounded, X closed, convex subset of X^* (i.e., an X compact subset of X^* .) Then K^* has continuous (in the X topology of X^*) nonzero, tangent functionals. In fact, it is known that these exist at each point of a dense subset of its boundary (see [7, Exercise 13, p. 459]). Explicitly, let D^* be the (nonempty) subset of the boundary of K^* at which continuous tangent functionals exist, i.e., for each $x_1^* \in D^*$ there exist a nonzero $g \in X$ and real constant c_g such that $g(K^*) \leq c_g$, $g(x_1^*) = c_g$. Such a g determines a support hyperplane h_g to K^* at x_1^* , where $h_g = \{x^* \in X : g(x^*) = c_g\}$, and a corresponding closed half-space, $H_g = \{x^* \in X^* : g(x^*) \leq c_g\}$, which contains K^* . Let G be the family of continuous tangent functionals so determined by elements of D^* .

LEMMA. K^* is uniquely determined as the intersection of the half-spaces H_g , i.e., $K^* = \bigcap_{g \in G} H_g$.

Proof. If $x^* \in K^*$ then $x^* \in H_g$ for every g , hence $K^* \subset \bigcap_{g \in G} H_g$.

To obtain the reverse inclusion, suppose $x_1^* \in \bigcap_{g \in G} H_g$ but $x_1^* \notin K^*$. Since K^* is closed and convex, there exists a continuous linear functional $x \in X$ which separates x_1^* and K^* . Suppose $x(x_1^*) = c$, $x(x^*) < c$, for $x^* \in K^*$. Let $c_x = \sup\{x(x^*) : x^* \in K^*\}$. Since K^* is compact in the X topology of X^* , $c_x < c$, and there exists an $x_2^* \in K^*$ such that $x(x_2^*) = c_x$. But then $x_2^* \in D^*$ and $x \in F$, and since $x(x_1^*) = c > c_x$, we have a contradiction to $x_1^* \in \bigcap_{g \in G} H_g$.

Remark. The existence of even a single support plane for a bounded, closed, convex subset of a Banach space is still an open question (see [10, p. 98]).

We now continue the proof of VIII. Let Ω_1 be the bounded, closed (in \mathcal{L}_∞) convex set which is not weak* closed, as shown to exist in VII. Let $\bar{\Omega}_1$ denote the weak* closure of Ω_1 ; then $\bar{\Omega}_1 - \Omega_1$ is not empty. Applying the preceding lemma to $\bar{\Omega}_1$, we see it is uniquely determined by its support planes; since $\Omega_1 \neq \bar{\Omega}_1$ there must be a support plane P to $\bar{\Omega}_1$ which is not a support plane of Ω_1 . Let $y \in \mathcal{L}_1$ be the continuous, linear, (tangent) functional which determines P , i.e., $y(x^*) \leq c$ for $x^* \in \bar{\Omega}_1$, and $y(x_1^*) = c$ for some $x_1^* \in \bar{\Omega}_1$. Since P is not a support plane for Ω_1 , $y(x^*) < c$ for all x^*

$\in \Omega_1$; but since x_1^* is in the weak* closure of Ω_1 , there exists a net $\{z_\nu^*\} \subset \Omega_1$ with $y(z_\nu^*)$ converging to c . This shows c is in the closure of $L(y)\Omega_1$ but not in $L(y)\Omega_1$.

Remark. Using the theorem of Lyapunov on the range of a vector measure, one can show there do exist closed, noncompact, subsets of \mathfrak{L}_∞ , e.g., $\{u \in \mathfrak{L}_\infty[0, T]: |u(t)| = 1\}$, which have the property that their image under any map of the form $L(y)$ is compact (see, for example, [4, Theorem 3] or [11, Theorem 1]).

IX. For simplicity we consider the interval $[-\pi, \pi]$ rather than $[0, T]$ and define the map $f: [-\pi, \pi] \rightarrow l_\infty$ with components f_i , $i = 0, 1, \dots$, as follows:

$$\begin{aligned} f_{2j}(t) &= \sin jt, & j &= 0, 1, \dots, \\ f_{2j+1}(t) &= \cos jt, & j &= 0, 1, \dots. \end{aligned}$$

Again, letting Σ denote the σ -algebra of Lebesgue measurable subsets of $[-\pi, \pi]$, we shall show that $S = \left\{ \int_{-\pi}^{\pi} f(\tau) \chi_E(\tau) d\tau: E \in \Sigma \right\}$ is not convex.

For $E \in \Sigma$, we may consider $\chi_E \in \mathfrak{L}_\infty[-\pi, \pi]$. Then since \mathfrak{L}_∞ is the dual of \mathfrak{L}_1 , there is a natural imbedding of \mathfrak{L}_1 into the dual of \mathfrak{L}_∞ , denoted \mathfrak{L}_∞^* . We shall identify the components f_i of f with their image under this imbedding, i.e., consider $\{f_i\}$ as a subset of \mathfrak{L}_∞^* . Then, from classical Fourier theory, $\text{span } \{f_i\}$ is total or, in other words, if $\int_{-\pi}^{\pi} u(\tau) f_i(\tau) d\tau = 0$ for all i and $u \in \mathfrak{L}_\infty[-\pi, \pi]$, then $u = 0$. Now the empty set $\emptyset \in \Sigma$ and $[-\pi, \pi] \in \Sigma$, therefore 0 and $\int_{-\pi}^{\pi} f d\tau$ are in S . If S were convex we would need $S_1 = \frac{1}{2} \int_{-\pi}^{\pi} f d\tau = \int_{-\pi}^{\pi} f(\tau) \frac{1}{2} d\tau$ in S . But since the span of the components f_i is total, there can be no element in \mathfrak{L}_∞ of the form χ_E , for $E \in \Sigma$ with $\int_{-\pi}^{\pi} f(\tau) \chi_E(\tau) d\tau = \int_{-\pi}^{\pi} f(\tau) \frac{1}{2} d\tau$, since the function u with $u(t) \equiv \frac{1}{2}$ is the unique element of \mathfrak{L}_∞ such that $\int_{-\pi}^{\pi} f u d\tau = s_1$.

3. Examples. (a) Any linear system of the form (4) with Ω as given in (i) and $U(t)$ convex and compact for each $t \in [0, T]$ can be transformed into an equivalent system which satisfies the hypotheses of V.

Indeed, let $X(t)$, $X(0) = I$, be a fundamental solution of the homogeneous system and make the change of variable $y(t) = X^{-1}(t)x(t)$. Then x satisfies (4) if and only if y satisfies $\dot{y}(t) = X^{-1}(t)B(t)u(t)$, $y(0) = x^0$. This transformed system obviously satisfies the hypotheses of V. Therefore, as is well known, the associated set $\mathfrak{A}(t)$ is compact and convex.

(b) Consider

$$\begin{aligned}\dot{x}_1 &= 1 + \sin x_2 u, & x_1(0) &= \pi, \\ \dot{x}_2 &= 1 - \sin x_2 u, & x_2(0) &= \pi, & 0 \leq u(t) \leq 2.\end{aligned}$$

Since $\dot{x}_1 \geq 0$, $\dot{x}_2 \geq 0$, $x \in \mathcal{A}(t)$ implies $x_1 \geq \pi$, $x_2 \geq \pi$, therefore $\{(f(x, u) : u \in U\}$ is independent of $x \in \mathcal{A}(t)$. (It is the segment of the line $y_1 + y_2 = 2$ with $y_1 \geq 0$, $y_2 \geq 0$.) The hypotheses of VI are satisfied and the attainable set will be compact and convex.

REFERENCES

- [1] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76–84.
- [2] J. P. LASALLE, *The time optimal control problem*, Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.
- [3] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110–117.
- [4] R. J. AUMANN, *Integrals of set valued functions*, Ibid., 12 (1965), pp. 1–12.
- [5] V. L. KLEE, JR., *Convex sets in linear spaces II*, Duke Math. J., 18 (1951), pp. 875–883.
- [6] H. HERMES, *Attainable sets and generalized geodesic spheres*, J. Differential Equations, 3 (1967), pp. 256–270.
- [7] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators I*, Interscience, New York, 1958.
- [8] P. L. FALB, *Infinite dimensional control problems I: on the closure of the set of attainable states for linear systems*, J. Math. Anal. Appl., 9 (1964), pp. 12–22.
- [9] YU. V. EGOROV, *Necessary conditions for the optimality of control in Banach spaces*, Mat. Sb. (N.S.), 64 (106) (1964), pp. 79–101.
- [10] V. L. KLEE, JR., *Extremal structure of convex sets II*, Math. Z., 69 (1958), pp. 90–104.
- [11] H. HERMES, *A note on the range of a vector measure; application to the theory of optimal control*, J. Math. Anal. Appl., 8 (1964), pp. 78–83.

SOME EXISTENCE THEOREMS FOR LINEAR OPTIMAL CONTROL PROBLEMS*

MARC Q. JACOBS†

1. Introduction. In this paper some existence theorems for linear optimal control problems with nonlinear cost functionals are presented. The aim of this paper is to take advantage of the strong assumptions concerning the linearity of the control system in order to prove some existence theorems for a class of cost functionals of the form $\int_0^{t_1} k(t, x, u) dt$, where k is lower semicontinuous and satisfies various convexity assumptions. An application of our results is given in §5. A comparison of our results with known results in the field is given in §3.

2. Preliminaries: Upper and lower semicontinuity in the space of closed subsets of R^q . Let $C(R^q)$ denote the collection of closed subsets of R^q . Then $C(R^q)$ is a complete lattice when ordered by set inclusion. As such, *upper* (*lower*) *semicontinuous* functions from an arbitrary topological space into $C(R^q)$ can be defined [19, pp. 73–74]. On the other hand, there is the concept of *upper* (*lower*) *semicontinuity with respect to inclusion*—in this context this is something of a misnomer—which is commonly used by control theorists, e.g., [9, p. 76]. The former notion of upper semicontinuity completely subsumes the latter, and is, in fact, precisely the property required of mappings into $C(R^q)$ in control theory investigations. For the sake of completeness, the principal results concerning such mappings are summarized below.

Let V be a subset of $R^q \times R^q$, and let A be a subset of R^q . Then we define

$$V[A] = \{x \in R^q \mid \exists a \in A :: (a, x) \in V\}.$$

Let d be any metric on R^q which is equivalent to the usual Euclidean metric. Denote the set $\{(x, y) \in R^q \times R^q \mid d(x, y) < \epsilon\}$ by J_ϵ^d , where $\epsilon > 0$. When the metric d is understood, then the family $\{J_\epsilon^d \mid \epsilon > 0\}$ will simply

* Received by the editors June 10, 1966, and in revised form December 13, 1966.

† University of Oklahoma, Norman, Oklahoma. Now at Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This paper is a portion of a dissertation submitted in partial fulfillment of the requirements of the Ph.D. degree in Mathematics at the University of Oklahoma. This research was supported in part by the Office of Aerospace Research of the Air Force Office of Scientific Research, The United States Air Force, under Grant AF-AFOSR-211-63, by a National Aeronautics and Space Administration Fellowship, the Office of Aerospace Research of the Air Force Office of Scientific Research, The United States Air Force, under Grant AF-AFOSR-693-66, and in part by the United States Army Research Office, Durham, under Contract DA-31-124-ARO-D-270.

be denoted by $\{J_\epsilon \mid \epsilon > 0\}$. If d is one of the *bounded* metrics on R^q equivalent to the usual Euclidean metric on R^q , then the Hausdorff metric [1, p. 111 ff.], [10, p. 166 ff.] d_H on $C(R^q) \setminus \{\emptyset\}$ is defined by

$$d_H(A, B) = \inf \{ \epsilon > 0 \mid J_\epsilon^d[A] \supset B \text{ and } J_\epsilon^d[B] \supset A \},$$

where $A, B \in C(R^q) \setminus \{\emptyset\}$.

A mapping \mathfrak{J} from a metric space S into $C(R^q)$ is *upper* (resp., *lower*) *semicontinuous with respect to inclusion* at $p_0 \in S$ (abbreviated: *usci* at $p_0 \in S$ (resp., *lsci* at $p_0 \in S$)) if and only if for every $\epsilon > 0$ there is a neighborhood $U(p_0)$ of p_0 such that $p \in U(p_0)$ implies $J_\epsilon[\mathfrak{J}(p_0)] \supset \mathfrak{J}(p)$ (resp., $J_\epsilon[\mathfrak{J}(p)] \supset \mathfrak{J}(p_0)$). We say that \mathfrak{J} is *upper* (resp., *lower*) *semicontinuous at* $p_0 \in S$ (abbreviated: *usc* at $p_0 \in S$ (resp., *lsc* at $p_0 \in S$)) if and only if $\limsup_{p \rightarrow p_0} \mathfrak{J}(p) \leq \mathfrak{J}(p_0)$ (resp., $\liminf_{p \rightarrow p_0} \mathfrak{J}(p) \geq \mathfrak{J}(p_0)$). The supremums and infimums are taken in $C(R^q)$ ordered by set inclusion. The definitions of upper and lower semicontinuity with respect to inclusion have been formulated in much greater generality by Berge [4, p. 109 ff.]. Similarly the definitions of upper and lower semicontinuity may be defined as long as the range of the function is in a complete lattice and the domain of the function is a topological space, e.g., see McShane and Botts [19, pp. 73-74]. We list below some of the consequences of these definitions which are useful in optimal control theory in general and in this paper in particular. The proofs are omitted. These and other results can be found in much greater generality in the author's dissertation.

THEOREM 1.1. *Let \mathfrak{J} be a mapping, $\mathfrak{J}: S \rightarrow C(R^q) \setminus \{\emptyset\}$, where S is a metric space. Then \mathfrak{J} is continuous with respect to the Hausdorff metric d_H on $C(R^q) \setminus \{\emptyset\}$ if and only if \mathfrak{J} is *usci* and *lsci* on S .*

THEOREM 1.2. *Let \mathfrak{J} be a mapping, $\mathfrak{J}: S \rightarrow C(R^q)$, where S is a locally compact metric space. Then the following are equivalent:*

- (i) \mathfrak{J} is *usc* at $p_0 \in S$;
- (ii) if $\{p_n\}, \{x_n\}$ are sequences in S and R^q respectively such that $x_n \in \mathfrak{J}(p_n)$, $n = 1, 2, 3, \dots$, and such that $p_n \rightarrow p_0$, $x_n \rightarrow x_0$ as $n \rightarrow \infty$, then $x_0 \in \mathfrak{J}(p_0)$.

THEOREM 1.3. *Let \mathfrak{J} be a mapping, $\mathfrak{J}: S \rightarrow C(R^q)$, where S is a metric space. If \mathfrak{J} is *usci* at $p_0 \in S$, then \mathfrak{J} is *usc* at $p_0 \in S$.*

THEOREM 1.4. *Let \mathfrak{J} be a mapping, $\mathfrak{J}: S \rightarrow C(R^q)$, where S is a metric space. Let $\mathfrak{J}(S) = \bigcup_{p \in S} \mathfrak{J}(p)$ be bounded. Then \mathfrak{J} is *usci* at $p_0 \in S$ if and only if \mathfrak{J} is *usc* at $p_0 \in S$.*

No use will be made of *lsc* or *lsci*, but in case there is interest, the following theorem is mentioned.

THEOREM 1.5. *Subject to the hypotheses of Theorem 1.4, we have that if \mathfrak{J} is *lsc* at $p_0 \in S$, then \mathfrak{J} is *lsci* at $p_0 \in S$.*

The following example shows that even when $\mathfrak{J}(p)$ is compact for each $p \in S$ we cannot infer that \mathfrak{J} being *usc* at p_0 implies \mathfrak{J} is *usci* at p_0 . Let the

mapping $\mathfrak{J}: [0, 1] \rightarrow C(R^2)$ be defined by

$$\mathfrak{J}(t) = \begin{cases} \{(0, 0)\} & \text{if } t = 0, \\ SQ(n) & \text{if } t = 1/n, \quad n = 1, 2, 3, \dots, \\ \{(n-1, n-1)\} & \text{if } 1/n < t < 1/(n-1), \quad n = 2, 3, 4, \dots, \end{cases}$$

where $SQ(n)$ is the boundary of the unit square with vertices $(n-1, n-1)$, $(n, n-1)$, $(n-1, n)$, (n, n) , $n = 1, 2, 3, \dots$. The function \mathfrak{J} is usc on $[0, 1]$, but \mathfrak{J} is not usc at $t = 0$.

3. Formulation of the linear optimal control problem. It is assumed that the control system at any time t can be described by a system of real ordinary differential equations of the following type,

$$(3.1) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

where for each $t \in [0, T]$ ($T > 0$ is fixed) $x(t)$ is an $n \times 1$ real matrix; $A(t)$ is an $n \times n$ real matrix; $B(t)$ is an $n \times m$ real matrix; $u(t)$ is an $m \times 1$ real matrix. No notational device will be used to distinguish row and column vectors; the context makes it clear which is meant. The matrices A and B in (3.1) are continuous on $[0, T]$. If x_0 is in R^n , and if u is a Lebesgue integrable function, $u: [0, t_1] \rightarrow R^m$, $t_1 \in [0, T]$, then there is a unique absolutely continuous function (response) $x(\cdot, u): [0, t_1] \rightarrow R^n$ satisfying (3.1) a.e. on $[0, t_1]$, and the initial condition

$$(3.2) \quad x(0, u) = x_0$$

(see [8, p. 74 ff.]). In fact the method of variation of parameters [8, p. 74 ff.] gives that

$$(3.3) \quad x(t, u) = X(t) \left[x_0 + \int_0^t X^{-1}(s)B(s)u(s) ds \right]$$

for $0 \leq t \leq t_1$, where the $n \times n$ matrix function X is defined by one matrix differential equation

$$\dot{X}(t) = A(t)X(t), \quad t \geq 0, \quad X(0) = I_n,$$

and I_n is the $n \times n$ identity matrix.

Let Ω denote a usc mapping, $\Omega: [0, T] \times R^n \rightarrow C(R^m)$, such that $\Omega(t, x)$ is nonempty, compact and convex for each (t, x) in $[0, T] \times R^n$. Denote by $S_0^t(\Omega)$, $0 \leq t \leq T$, the collection $\{u \mid u: [0, t] \rightarrow R^m, u \text{ measurable on } [0, t], u(s) \in \Omega(s, x(s, u)), \forall s \in [0, t]\}$; this set may be empty. Let Γ be a nonempty closed subset of $[0, T]$. Define a set $U(\Omega, \Gamma)$ by the relation

$$(3.4) \quad U(\Omega, \Gamma) = \bigcup_{t \in \Gamma} S_0^t(\Omega) \times \{t\}.$$

Let a target function $F: \Gamma \rightarrow C(R^n)$ be given such that $F(t)$ is nonempty

for each $t \in \Gamma$. The function F is required to be usc on Γ . The reader is reminded that this is a strictly weaker requirement than asking that F be usc on Γ . Define the set $\hat{U}(\Omega, \Gamma)$ to be the subset of $U(\Omega, \Gamma)$ defined by the condition that

$$(3.5) \quad (u, t_1) \in \hat{U}(\Omega, \Gamma) \quad \text{if and only if} \\ (u, t_1) \in \mathcal{S}_0^{t_1}(\Omega) \times \{t_1\} \quad \text{and} \quad x(t_1, u) \in F(t_1).$$

An element (u, t_1) in $\hat{U}(\Omega, \Gamma)$, with u defined on $[0, t_1]$ and $t_1 \in \Gamma$, is said to be an admissible control function, or an admissible steering function, or simply an admissible control.

We shall now consider a scalar function k , defined on $[0, T] \times R^n \times R^m$, for which we assume that

$$(3.6) \quad \begin{aligned} &k: [0, T] \times R^n \times R^m \rightarrow R \text{ is a lower semicontinuous (lsc)} \\ &\text{function of } (t, x, u) \text{ satisfying } |k(t, x, u)| \leq \mu(t)g(\|x\|), \\ &\text{where } \mu \text{ is summable on } [0, T], \text{ and } g(\|x\|) = O(\|x\|) \\ &\text{as } \|x\| \rightarrow \infty. \end{aligned}$$

Using the mapping k we shall now define the function $K: U(\Omega, \Gamma) \rightarrow R$ by taking

$$(3.7) \quad K(u, t_1) = \int_0^{t_1} k(s, x(s, u), u(s)) \, ds$$

when (u, t_1) is in $U(\Omega, \Gamma)$. The existence of the integral in (3.7) can be established by using a theorem in [18, p. 123]. The mapping K is termed the *cost functional* (the criterion of optimality or the objective). The optimal control problem studied in this paper is

$$(3.8) \quad K(u, t_1) = \text{minimum on } \hat{U}(\Omega, \Gamma).$$

The results of Lee and Markus [15] can be applied to (3.8) if the function k is of the form

$$k(t, x, u) = a(t, x) + \sum_{i=1}^m b_i(t, x)u^i,$$

and a, b_i are continuous, $i = 1, 2, \dots, m$, Ω is a fixed compact convex subset of R^m , $\Gamma = [0, T]$, $F(t)$ is compact for each t in $[0, T]$, the mapping F is continuous with respect to the Hausdorff metric, and $\hat{U}(\Omega, \Gamma)$ is nonempty. Lee and Markus permit some nonlinearity in the control system (3.1), but if the control system is linear as in (3.1), then Lee and Markus' result falls as a corollary to our more general Theorem 4.1 or 4.3. In [14] the time-optimal control problem ($k = 1$, Ω the unit cube in R^m) was

studied in great detail. Existence theorems for the time-optimal linear control problem are simple consequences of our results. Balakrishnan [2] studied the problem of minimizing $\|x(T, u) - x_1\|$ on \mathfrak{B} , a closed and bounded convex subset of $L_2^m[0, T]$, x_1 a fixed point in R^n . The existence theorems of Roxin [24] and Filippov [9] cover (3.8), if for each fixed $(t, x) \in [0, T] \times R^n$, $H(t, x, \Omega)$ is a compact convex subset of R^{n+1} , where

$$H(t, x, u) = (A(t)x + B(t)u, k(t, x, u)), \quad u \in \Omega,$$

and Ω is a fixed compact subset of R^m . It is also assumed that k is continuous in (x, u) for each fixed t , k is summable with respect to t for each fixed (x, u) , k is Lipschitzian in x , and k satisfies (3.6). These results and ours speak to an overlapping class of problems, but neither contains the other as a special case. Roxin and Filippov allow for some nonlinearity in the control system, whereas we do not. Neustadt [21] removes all assumptions of convexity from both the control system (3.1) and the restraint set Ω , but his existence theorem covers (3.8) only if the mapping k in (3.6) has the form

$$k(t, x, u) = \sum_{i=1}^n \alpha_i(t) x^i + \phi^0(u, t),$$

and $\alpha_i : [0, T] \rightarrow R$, $\phi^0 : [0, T] \times R^m \rightarrow R$ are continuous, $i = 1, 2, \dots, m$. Neither our results nor Neustadt's necessarily include the other. Chang [7] assumes the matrices A and B in (3.1) are constant, and that the cost functional in (3.7) has the form

$$K(u) = \int_0^\infty \left\{ \sum_{i,j=1}^n q_{ij} x^i(t, u) x^j(t, u) + \sum_{i,j=1}^m r_{ij} u^i(t) u^j(t) \right\} dt,$$

where (q_{ij}) is a nonnegative definite $n \times n$ matrix of real constants, (r_{ij}) is an $m \times m$ positive definite matrix of real constants. The restraint set is the "unit cube" in R^m , and the domain of each control is $[0, \infty]$. With these hypotheses Chang has used the Banach-Saks theorem [23, p. 80] to prove there exists an admissible control (i.e., a measurable function with range in the unit cube of R^m which transfers an initial point x_1 in R^n to the origin as $t \rightarrow \infty$) for which the function K achieves its minimum. Cesari's analysis [4], [5] concerns differential systems and integrand functions k which in general are nonlinear. One of Cesari's theorems [5, Theorem I, p. 478], when applied to the situation considered here, yields a statement which is partially overlapping with ours in the sense that Cesari assumes the map k only to be continuous in (t, x, u) and convex in u , while in our statement (Theorem 4.1 below) k is lsc in (t, x, u) , Lipschitzian in x , and convex in u . This result of Cesari's, however, requiring the continuity of f , cannot be applied to the example which we give in §5, whereas our Theo-

rems 4.2 and 4.3 (below) can both be applied to this example. It should also be pointed out that if u enters in the control system in a nonlinear fashion—say, $(\dot{x}^1, \dot{x}^2) = (\phi^1(u^1, u^2), \phi^2(u^1, u^2))$, where ϕ^1 and ϕ^2 are third degree polynomials in u^1, u^2 —then Cesari [4, p. 13] shows by means of a counterexample that our results can no longer be expected to be true, even if k is continuous.

4. Existence theorems for the linear optimal control problem within the class $\hat{U}(\Omega, \Gamma)$. If $\{u_p\}$ is a sequence in $L_2^m[a, b]$ which converges weakly to u in $L_2^m[a, b]$, then we shall write

$$u_p \rightarrow u \text{ (wk) as } p \rightarrow \infty,$$

and if a sequence $\{u_p\}$ in $L_2^m[a, b]$ converges to u in $L_2^m[a, b]$ with respect to the inner product norm (strongly), then we shall write

$$u_p \rightarrow u \text{ (st) as } p \rightarrow \infty.$$

We now enumerate the collection of hypotheses that will be used in this section:

- (H₁) (3.6) is satisfied;
- (H₂) the matrix-valued functions A and B appearing in (3.1) are continuous on $[0, T]$;
- (H₃) Γ is a nonempty closed subset of $[0, T]$;
- (H₄) the mapping $\Omega: [0, T] \times R^n \rightarrow C(R^m)$ is usc_i, and $\Omega(t, x)$ is compact, convex and nonempty for each $(t, x) \in [0, T] \times R^n$, and moreover, the set Z defined by $Z = \bigcup_{(t,x) \in [0,T] \times R^n} \Omega(t, x)$ is bounded;
- (H₅) the mapping $F: \Gamma \rightarrow C(R^n)$ is usc, and $F(t)$ is nonempty for each $t \in \Gamma$;
- (H₆) the function k in (3.6) is convex in u , i.e., for each fixed $(t, x) \in [0, T] \times R^n$ we have that

$$k\left(t, x, \frac{u+v}{2}\right) \leq \frac{1}{2} k(t, x, u) + \frac{1}{2} k(t, x, v), \quad u, v \in R^m;$$

- (H₆') the function k in (3.6) is convex in (x, u) , i.e., for each fixed $t \in [0, T]$ we have that

$$k\left(t, \frac{x+y}{2}, \frac{u+v}{2}\right) \leq \frac{1}{2} k(t, x, u) + \frac{1}{2} k(t, y, v), \quad x, y \in R^n, \quad u, v \in R^m;$$

(H₇) k satisfies a Lipschitz condition in the variable x , i.e., there is a constant $A > 0$ such that for each fixed (t, u) in $[0, T] \times [\bigcup_{(t,x) \in \Gamma \times \mathbb{R}^n} \Omega(t, x)]$ we have that

$$|k(t, x, u) - k(t, y, u)| \leq A |x - y|, \quad x, y \in \mathbb{R}^n;$$

(H₈) $\mathcal{S}_0^t(\Omega)$ is nonempty for $t \in \Gamma$;

(H₉) $\hat{U}(\Omega, \Gamma)$ is nonempty.

In connection with (H₈), it should be pointed out that if the mapping Ω is independent of x , then (H₈) follows from the assumption that Ω is usc on $[0, T]$. Hypothesis (H₈) is employed to insure that the system of differential equations (3.1) and the restraint set $\Omega(t, x)$ are compatible.

DEFINITION 4.1. $\hat{U}(\Omega, \Gamma)$ is weakly compact in itself if and only if for any sequence $\{(u_p, t_p)\}$ in $\hat{U}(\Omega, \Gamma)$ there is a (u_0, t_0) in $\hat{U}(\Omega, \Gamma)$, such that some subsequence $\{u_{p_j}, t_{p_j}\}$ of $\{(u_p, t_p)\}$ has the property that

$$\tilde{u}_{p_j} \rightarrow u_0 \text{ (wk)}, \quad t_{p_j} \rightarrow t_0 \quad \text{as } j \rightarrow \infty,$$

where \tilde{u}_{p_j} in $L_2^m[0, t_0]$ is given by $\tilde{u}_{p_j} = u_{p_j} \mid [0, t_0]$ if $t_0 \leq t_{p_j}$ ($u_{p_j} \mid [0, t_0]$ denotes the restriction of u_{p_j} to $[0, t_0]$), and if $t_0 > t_{p_j}$ the function \tilde{u}_{p_j} is given by

$$\tilde{u}_{p_j}(t) = \begin{cases} u_{p_j}(t) & \text{if } 0 \leq t \leq t_{p_j}, \\ u_0(t) & \text{if } t_{p_j} < t \leq t_0. \end{cases}$$

We now have the lemma.

LEMMA 4.1. If (H₂), (H₃), (H₄), (H₅) and (H₈) are satisfied, then $\hat{U}(\Omega, \Gamma)$ is either empty or weakly compact in itself.

Proof. Let $\{(u_p, t_p)\}$ be a sequence of elements in $\hat{U}(\Omega, \Gamma)$. It may be assumed that $t_p \rightarrow t_0 \in \Gamma$ (monotonely) as $p \rightarrow \infty$. There are two cases:

Case 1. $0 \leq t_0 \leq t_p \leq T$, $p = 1, 2, 3, \dots$.

Case 2. $0 \leq t_p \leq t_0 \leq T$, $p = 1, 2, 3, \dots$.

We deal with Case 1 first. It follows from (H₄) that the functions \tilde{u}_p are uniformly bounded with respect to the norm of $L_2^m[0, t_0]$. Consequently, there is a subsequence of $\{(u_p, t_p)\}$ (still denoted by $\{(u_p, t_p)\}$) such that $\tilde{u}_p \rightarrow u_0$ (wk) as $p \rightarrow \infty$ [17, p. 117]. Thus the subsequences $\{\tilde{u}_{p_N}\} = \{\tilde{u}_p, p \geq N\}$, $N = 1, 2, 3, \dots$, each converge weakly to u_0 . By the Banach-Saks theorem [23, p. 80] and a theorem relating strong convergence to convergence almost everywhere [13, p. 87], we obtain that for each $N = 1, 2, 3, \dots$ there is a sequence $\{\sigma_{p_N}\}$ of convex linear combinations of the \tilde{u}_{p_N} , such that $\sigma_{p_N}(t) \rightarrow u_0(t)$ as $p \rightarrow \infty$ for almost every $t \in [0, t_0]$. Let E_N denote the set $\{t \in [0, t_0] \mid \lim_{p \rightarrow \infty} \sigma_{p_N}(t) \neq u_0(t)\}$, $N = 1, 2, 3, \dots$. Then each E_N has measure zero, and thus $E = \bigcup_{N=1}^{\infty} E_N$ also has measure zero. From (3.3) and the fact that $\tilde{u}_p \rightarrow u_0$ (wk) as $p \rightarrow \infty$ it follows that $x(t, u_p) \rightarrow x(t, u_0)$ as $p \rightarrow \infty$, $0 \leq t \leq t_0$. If $t \in [0, t_0] \setminus E$, and if $\epsilon > 0$ is

given, then it follows from (H_4) that there is a positive integer $N(\epsilon, t)$ such that

$$p \geq N(\epsilon, t) \quad \text{implies} \quad J_\epsilon[\Omega(t, x(t, u_0))] \supset \Omega(t, x(t, u_p));$$

any norm on R^m can be used to define the sets J_ϵ , $\epsilon > 0$. Now $\tilde{u}_{pN(\epsilon, t)} = u_p$ for $p \geq N(\epsilon, t)$, and therefore we have that $\tilde{u}_{pN(\epsilon, t)}(t) \in \Omega(t, x(t, u_p))$ for $p \geq N(\epsilon, t)$. The set $J_\epsilon[\Omega(t, x(t, u_0))]$ is convex. Thus it follows that $\sigma_{pN(\epsilon, t)}(t) \in J_\epsilon[\Omega(t, x(t, u_0))]$ for $p \geq N(\epsilon, t_0)$. Consequently $u_0(t)$ is a point of the closure of $J_\epsilon[\Omega(t, x(t, u_0))]$; and since $\Omega(t, x(t, u_0))$ is compact and $\epsilon > 0$ is arbitrary, we have thereby shown $u_0(t) \in \Omega(t, x(t, u_0))$ for the given t in $[0, t_0] \setminus E$. By suitably redefining u_0 on the exceptional set E of measure zero the function u_0 has the property that $u_0(t) \in \Omega(t, x(t, u_0))$, $0 \leq t \leq t_0$. We also have from the definition of $\hat{U}(\Omega, \Gamma)$ that $x(t_p, u_p) \in F(t_p)$, $p = 1, 2, 3, \dots$. A straightforward calculation reveals that

$$\begin{aligned} \lim_{p \rightarrow \infty} x(t_p, u_p) &= \lim_{p \rightarrow \infty} [x(t_p, u_p) - x(t_p, u_p) + x(t_0, u_p)] \\ &= \lim_{p \rightarrow \infty} x(t_0, u_p) = x(t_0, u_0). \end{aligned}$$

Theorem 1.3 applies to give $x(t_0, u_0) \in F(t_0)$, thereby proving $(u_0, t_0) \in \hat{U}(\Omega, \Gamma)$. In the event that Case 2 obtains, each control is extended to the entire interval $[0, t_0]$ in the following way:

$$\tilde{u}_p(t) = \begin{cases} u_p(t) & \text{if } 0 \leq t \leq t_p, \\ u^*(t) & \text{if } t_p \leq t \leq t_0, \end{cases}$$

where u^* is any function in $S_0^{t_0}(\Omega)$ which is nonempty by (H_3) . A subsequence of $\{\tilde{u}_p\}$ (still denoted by $\{\tilde{u}_p\}$) converges weakly to a $u_0 \in L_2^m[0, t_0]$. By slight modifications of the proof for Case 1, it can be shown that $\tilde{u}_p \rightarrow u_0$ (wk), $t_p \rightarrow t_0$ as $p \rightarrow \infty$ and that $(u_0, t_0) \in \hat{U}(\Omega, \Gamma)$.

THEOREM 4.1. *If (H_1) , (H_2) , (H_3) , (H_4) , (H_5) , (H_6) , (H_7) , (H_8) and (H_9) are satisfied, then K has an absolute minimum on $\hat{U}(\Omega, \Gamma)$, i.e., there is a (u_0, t_0) in $\hat{U}(\Omega, \Gamma)$ such that*

$$K(u_0, t_0) = \inf K(\hat{U}(\Omega, \Gamma)).$$

Proof. We first observe that

$$\mathcal{B} = \{x \mid x = x(t, u), (u, t_1) \in \hat{U}(\Omega, \Gamma), 0 \leq t \leq t_1\}$$

is a bounded subset of R^n . This follows immediately from (3.3), if one notices that X , X^{-1} and B are all continuous on $[0, T]$, and that the set $Z = \bigcup_{(t, x) \in [0, T] \times R^n} \Omega(t, x)$ is bounded. Consequently since k is lsc on $[0, T] \times R^n \times R^m$, it follows that the set of real numbers $k([0, T] \times \mathcal{B} \times Z)$ is bounded from below. Therefore, by (H_9) , $K(\hat{U}(\Omega, \Gamma))$ is a non-empty set of real numbers bounded from below. Whence, if γ denotes

the infimum of $K(\hat{U}(\Omega, \Gamma))$, then $+\infty > \gamma > -\infty$. There is a “minimizing sequence” of controls

$$(4.1) \quad \{(u_p, t_p)\} \subset \hat{U}(\Omega, \Gamma), \quad K(u_p, t_p) \rightarrow \gamma \quad \text{as } p \rightarrow \infty.$$

Clearly this “minimizing sequence” admits a subsequence (still called $\{(u_p, t_p)\}$) such that $t_p \rightarrow$ some t_0 in Γ monotonely as $p \rightarrow \infty$. There are two cases:

Case 1. $0 \leq t_0 \leq t_p \leq T, p = 1, 2, 3, \dots$

Case 2. $0 \leq t_p \leq t_0 \leq T, p = 1, 2, 3, \dots$

Case 1 is considered first. In view of Lemma 4.1, there is a further subsequence of $\{(u_p, t_p)\}$ (still denoted by $\{(u_p, t_p)\}$) such that

$$(4.2) \quad u_p \rightarrow u_0 \text{ (wk)}, \quad t_p \rightarrow t_0 \quad \text{as } p \rightarrow \infty, \quad (u_0, t_0) \in \hat{U}(\Omega, \Gamma);$$

whence by the Banach-Saks theorem [23, p. 80] there is a further subsequence of $\{(u_p, t_p)\}$ (without changing the notation) such that

$$(4.3) \quad \sigma_p = \frac{1}{p} \sum_{i=1}^p u_i \rightarrow u_0 \text{ (st)}, \quad t_p \rightarrow t_0 \quad \text{as } p \rightarrow \infty.$$

Thus (see [13, p. 87]) there is a subsequence of $\{(\sigma_p, t_p)\}$, say $\{(\sigma_{p_j}, t_{p_j})\}$, such that

$$(4.4) \quad \sigma_{p_j} \rightarrow u_0 \text{ (a.e.) on } [0, t_0] \quad \text{as } j \rightarrow \infty.$$

From (3.3) and the fact that $u_p \rightarrow u_0$ (wk) as $p \rightarrow \infty$, we have that

$$(4.5) \quad x(t, u_p) \rightarrow x(t, u_0) \quad \text{as } p \rightarrow \infty, \quad 0 \leq t \leq t_0.$$

It thereby follows that

$$(4.6) \quad \lim_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} x(t, u_i) = \lim_{p \rightarrow \infty} x(t, u_p) = x(t, u_0), \quad 0 \leq t \leq t_0.$$

Since k is convex in u by (H_6) , we have that (see (4.3))

$$(4.7) \quad k(s, x(s, u_{p_j}), \sigma_{p_j}(s)) \leq \frac{1}{p_j} \sum_{i=1}^{p_j} k(s, x(s, u_{p_j}), u_i(s))$$

for $0 \leq s \leq t_0, j = 1, 2, 3, \dots$. Also the following equalities are valid:

$$(4.8a) \quad \lim_{p \rightarrow \infty} K(u_p, t_p) = \lim_{j \rightarrow \infty} K(u_{p_j}, t_{p_j}) = \gamma,$$

$$(4.8b) \quad \begin{aligned} \lim_{j \rightarrow \infty} K(u_{p_j}, t_{p_j}) &= \lim_{j \rightarrow \infty} \int_0^{t_{p_j}} k(s, x(s, u_{p_j}), u_{p_j}(s)) \, ds \\ &= \lim_{j \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_{p_j}), u_{p_j}(s)) \, ds \\ &= \lim_{p \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_p), u_p(s)) \, ds. \end{aligned}$$

Relation (4.8a) is evident, whereas (4.8b) is an easy consequence of (H_1) .

The next step in the proof is to show that

$$(4.9) \quad \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{\tau_0} k(s, x(s, u_{p_j}), u_i(s)) ds \rightarrow \gamma \quad \text{as } j \rightarrow \infty.$$

By (4.8a) and (4.8b), we obtain that

$$(4.10) \quad \lim_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{\tau_0} k(s, x(s, u_i), u_i(s)) ds = \gamma.$$

Thus in order to prove (4.9) it will suffice to prove that

$$(4.11) \quad \lim_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{\tau_0} [k(s, x(s, u_{p_j}), u_i(s)) - k(s, x(s, u_i), u_i(s))] ds = 0.$$

From (H₇) we obtain the inequality

$$(4.12) \quad \left| \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{\tau_0} [k(s, x(s, u_{p_j}), u_i(s)) - k(s, x(s, u_i), u_i(s))] ds \right| \leq \frac{1}{p_j} \sum_{i=1}^{p_j} A \int_0^{\tau_0} \|x(s, u_{p_j}) - x(s, u_i)\| ds.$$

We have that

$$\lim_{j \rightarrow \infty} x(s, u_{p_j}) = \lim_{i \rightarrow \infty} x(s, u_i) = x(s, u_0), \quad 0 \leq s \leq \tau_0,$$

and it follows from the Lebesgue dominated convergence theorem (or even the bounded convergence theorem) that

$$\lim_{i \rightarrow \infty} \int_0^{\tau_0} \|x(s, u_i) - x(s, u_0)\| ds = 0$$

and

$$\lim_{j \rightarrow \infty} \int_0^{\tau_0} \|x(s, u_{p_j}) - x(s, u_0)\| ds = 0.$$

Thus given $\epsilon > 0$ there is a positive integer p_ϵ^0 , which can be taken to be one of the integers in the collection $p_1 < p_2 < p_3 < \cdots < p_j < \cdots$, such that

$$(4.13) \quad p_j, i \geq p_\epsilon^0 \quad \text{implies} \quad A \int_0^{\tau_0} \|x(s, u_{p_j}) - x(s, u_i)\| ds < \epsilon/2.$$

This p_ϵ^0 is fixed (it depends only on the given $\epsilon > 0$). Now since $x(s, u_{p_j})$, $x(s, u_i)$, $i, j = 1, 2, 3, \cdots$, are uniformly bounded on $[0, \tau_0]$, it follows that there is a $\lambda > 0$, such that

$$(4.14) \quad 0 \leq \sum_{i=1}^{p_\epsilon^0} \int_0^{\tau_0} A \|x(s, u_{p_j}) - x(s, u_i)\| ds \leq \lambda$$

for all $p_j, j = 1, 2, 3, \dots$. There is an integer $p_\epsilon > p_\epsilon^0$ such that

$$(4.15) \quad p_j \geq p_\epsilon > p_\epsilon^0 \text{ implies } \lambda/p_j < \epsilon/2.$$

Denote the sum on the right-hand side of (4.12) by S_j . Then given $p_j \geq p_\epsilon$, it follows that

$$\begin{aligned} 0 \leq S_j &= \frac{1}{p_j} \sum_{i=1}^{p_\epsilon^0} \int_0^{t_0} A \|x(s, u_{p_j}) - x(s, u_i)\| ds \\ &\quad + \frac{1}{p_j} \sum_{i=p_\epsilon^0+1}^{p_j} \int_0^{t_0} A \|x(s, u_{p_j}) - x(s, u_i)\| ds. \end{aligned}$$

Consequently by (4.13), (4.14) and (4.15), one obtains that

$$(4.16) \quad p_j \geq p_\epsilon > p_\epsilon^0 \text{ implies } 0 \leq S_j < \epsilon/2 + \frac{1}{p_j} \sum_{i=p_\epsilon^0+1}^{p_j} [\epsilon/2] < \epsilon.$$

Therefore, $S_j \rightarrow 0$ as $j \rightarrow \infty$, and (4.11) is true in view of inequality (4.12). This establishes the validity of (4.9).

Returning to inequality (4.7), we recall that Z is bounded, $\sigma_{p_j}(t) \in Z$ for $0 \leq t \leq t_0 \leq T$, and the responses $x(\cdot, u_p)$, $p = 1, 2, 3, \dots$, are uniformly bounded on $[0, t_0]$. It follows then from the lower semicontinuity of k that there is a real number δ such that

$$(4.17) \quad k(s, x(s, u_{p_j}), \sigma_{p_j}(s)) \geq \delta, \quad 0 \leq s \leq t_0, \quad j = 1, 2, 3, \dots$$

From (H_1) , (4.4) and (4.5), we find that

$$(4.18) \quad \liminf_{j \rightarrow \infty} k(s, x(s, u_{p_j}), \sigma_{p_j}(s)) \geq k(s, x(s, u_0), u_0(s))$$

a.e. on $[0, t_0]$. Utilizing (4.7), we find that

$$\begin{aligned} (4.19) \quad \liminf_{j \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_{p_j}), \sigma_{p_j}(s)) ds \\ \leq \liminf_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{t_0} k(s, x(s, u_{p_j}), u_i(s)) ds. \end{aligned}$$

By virtue of (4.17), (4.18), Fatou's lemma [18, p. 167] can be applied to the left-hand side of (4.18) to obtain the relation

$$(4.20) \quad \liminf_{j \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_{p_j}), \sigma_{p_j}(s)) ds \geq \int_0^{t_0} k(s, x(s, u_0), u_0(s)) ds.$$

In view of the definition of K , relations (4.9), (4.19) and (4.20), we infer that

$$(4.21) \quad K(u_0, t_0) \leq \gamma.$$

But as a consequence of the definition of γ and the fact that (u_0, t_0) is

in $\hat{U}(\Omega, \Gamma)$, we also have that

$$(4.21') \quad K(u_0, t_0) \geq \gamma.$$

Combining (4.21) and (4.21') we deduce that $K(u_0, t_0) = \gamma$, thus completing the proof for Case 1.

In order to treat Case 2, we extend each control in the "minimizing subsequence" to the entire interval $[0, t_0]$ by choosing an element u^* of $S_0^{t_0}(\Omega)$, and defining

$$\bar{u}_p(t) = \begin{cases} u_p(t) & \text{if } 0 \leq t \leq t_p, \\ u^*(t) & \text{if } t_p < t \leq t_0, \end{cases}$$

for $p = 1, 2, 3, \dots$. Case 2 can then be disposed of by retracing the steps used in the proof of Case 1 but utilizing the "modified minimizing subsequence" in lieu of the one used in that proof.

A good many of the details in the proofs of Theorems 4.2 and 4.3 (below) are quite similar to those given in the proof of Theorem 4.1. For this reason we shall use an abbreviated exposition in the proofs of the next two theorems.

THEOREM 4.2. *If $(H_1), (H_2), (H_3), (H_4), (H_5), (H_6'), (H_8)$ and (H_9) are satisfied, then the functional K has an absolute minimum on $\hat{U}(\Omega, \Gamma)$.*

Proof. Denote by γ the infimum of the set $K(\hat{U}(\Omega, \Gamma))$. There is a sequence

$$(4.22) \quad (u_p, t_p) \in \hat{U}(\Omega, \Gamma), \quad p = 1, 2, 3, \dots,$$

such that

$$(4.23a) \quad t_p \rightarrow t_0 \in \Gamma \quad (\text{monotonely}) \quad \text{as } p \rightarrow \infty,$$

$$(4.23b) \quad u_p \rightarrow u_0 \text{ (wk) as } p \rightarrow \infty, \quad (u_0, t_0) \in \hat{U}(\Omega, \Gamma),$$

$$(4.23c) \quad K(u_p, t_p) \rightarrow \gamma > -\infty \quad \text{as } p \rightarrow \infty.$$

There are two cases:

Case 1. $0 \leq t_0 \leq t_p \leq T, p = 1, 2, 3, \dots$

Case 2. $0 \leq t_p \leq t_0 \leq T, p = 1, 2, 3, \dots$

Consider Case 1. By the Banach-Saks theorem [23, p. 80] there is a subsequence of $\{(u_p, t_p)\}$, which we still denote by $\{(u_p, t_p)\}$, such that

$$(4.24) \quad \sigma_p = \frac{1}{p} \sum_{i=1}^p u_i \rightarrow u_0 \text{ (st), } t_p \rightarrow t_0 \quad \text{as } p \rightarrow \infty.$$

Then (see [13, p. 87]) there is a subsequence of $\{(\sigma_p, t_p)\}$, say $\{(\sigma_{p_j}, t_{p_j})\}$, such that

$$(4.25) \quad \sigma_{p_j} \rightarrow u_0 \text{ (a.e.) on } [0, t_0], \quad t_{p_j} \rightarrow t_0 \quad \text{as } j \rightarrow \infty.$$

It follows from (4.23b) and (3.3) that

$$(4.26) \quad x(t, u_p) \rightarrow x(t, u_0) \quad \text{as } p \rightarrow \infty, \quad 0 \leq t \leq t_0.$$

From (4.26), (4.23b) and (4.23c) we obtain that

$$(4.27a) \quad \lim_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} x(t, u_i) = x(t, u_0), \quad 0 \leq t \leq t_0,$$

$$(4.27b) \quad \lim_{p \rightarrow \infty} K(u_p, t_p) = \gamma = \lim_{p \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_p), u_p(s)) \, ds,$$

$$(4.27c) \quad \lim_{j \rightarrow \infty} \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{t_0} k(s, x(s, u_i), u_i(s)) \, ds = \gamma.$$

The convexity condition (H_6) on k gives that

$$(4.28) \quad \int_0^{t_0} k\left(s, \frac{1}{p_j} \sum_{i=1}^{p_j} x(s, u_i), \frac{1}{p_j} \sum_{i=1}^{p_j} u_i(s)\right) \, ds \\ \leq \frac{1}{p_j} \sum_{i=1}^{p_j} \int_0^{t_0} k(s, x(s, u_i), u_i(s)) \, ds.$$

Then in a manner entirely similar to that presented in the proof of Theorem 4.1 one may establish that

$$\gamma \leq K(u_0, t_0) \leq \gamma.$$

This completes the proof for Case 1.

The details of the proof for Case 2 will be omitted.

The force of Theorem 4.2 is that the Lipschitz condition (hypothesis (H_7) of Theorem 4.1) on k may be dispensed with if we require that k be a convex function in both the variables x and u . The question which now arises is: Can the Lipschitz condition in the variable x and the convexity condition in the variable x both be omitted? Insofar as the author has been able to discover, the answer is a qualified yes, but at the expense of imposing other very restrictive hypotheses on the function k such as those given in the following theorem.¹

THEOREM 4.3. *Let (H_2) , (H_3) , (H_4) , (H_5) , (H_8) and (H_9) be satisfied and let k be of the form*

$$k(t, x, u) = a(t, x) + \sum_{i=1}^m b_i(t, x) u^i(t) + \sum_{i,j=1}^m c_{ij}(t, x) u^i(t) u^j(t),$$

where the mappings $a, b_i, c_{ij} : [0, T] \times R^n \rightarrow R$, $i, j = 1, 2, \dots, m$, are required to satisfy the following two conditions:

(i) $a, b_i, c_{ij}, i, j = 1, 2, \dots, m$, are each lsc in the variables $(t, x) \in [0, T] \times R^n$ and continuous in the variable $x \in R^n$, and, moreover, $|a(t, x)|$,

¹ Cesari's result [5, Theorem 1] provides another alternative if we assume k is continuous in (t, x, u) and convex in u .

$|b_i(t, x)|, |c_{ij}(t, x)| \leq \mu(t)g(\|x\|), i, j = 1, 2, \dots, m$, where μ and g are as in (H_1) ;

(ii) for each $(t, x, u) \in [0, T] \times R^n \times R^m$ we have that $\sum_{i,j=1}^m c_{ij}(t, x)u^i u^j \geq 0$, and $c_{ij} = c_{ji}, i, j = 1, 2, \dots, m$.

Then the functional K has an absolute minimum on $\hat{U}(\Omega, \Gamma)$.

Proof. In accordance with Lemma 4.1 there is a "minimizing sequence" $\{(u_p, t_p)\}$ in $\hat{U}(\Omega, \Gamma)$ such that

$$(4.29a) \quad u_p \rightarrow u_0 \text{ (wk)}, \quad t_p \rightarrow t_0 \text{ (monotonely) as } p \rightarrow \infty, \\ (u_0, t_0) \in \hat{U}(\Omega, \Gamma),$$

$$(4.29b) \quad K(u_p, t_p) \rightarrow \gamma > -\infty \text{ as } p \rightarrow \infty,$$

where γ denotes the infimum of the set $K(\hat{U}(\Omega, \Gamma))$.

We consider the two cases:

Case 1. $0 \leq t_0 \leq t_p \leq T, p = 1, 2, 3, \dots$

Case 2. $0 \leq t_p \leq t_0 \leq T, p = 1, 2, 3, \dots$

First assume that Case 1 holds. Then from (4.29a) and (3.3) it follows that

$$(4.30) \quad x(t, u_p) \rightarrow x(t, u_0) \text{ as } p \rightarrow \infty, \quad t \in [0, t_0].$$

By Theorem 4.3(i), (4.30) and the Lebesgue dominated convergence theorem (its applicability is easily justified), it is determined that

$$(4.31) \quad \lim_{p \rightarrow \infty} \int_0^{t_0} a(s, x(s, u_p)) ds = \int_0^{t_0} a(s, x(s, u_0)) ds.$$

Also by (4.30) and Theorem 4.3(i), we deduce that

$$(4.32) \quad \lim_{p \rightarrow \infty} b_i(t, x(t, u_p)) = b_i(t, x(t, u_0))$$

for $i = 1, 2, 3, \dots, m$, and $0 \leq t \leq t_0$. The following inequality is easily seen to be valid:

$$(4.33) \quad \left| \int_0^{t_0} \sum_{i=1}^m [b_i(s, x(s, u_p))u_p^i(s) - b_i(s, x(s, u_0))u_0^i(s)] ds \right| \\ \leq \left| \int_0^{t_0} \sum_{i=1}^m [b_i(s, x(s, u_p)) - b_i(s, x(s, u_0))]u_p^i(s) ds \right| \\ + \left| \int_0^{t_0} \sum_{i=1}^m b_i(s, x(s, u_0))[u_p^i(s) - u_0^i(s)] ds \right|$$

for $p = 1, 2, 3, \dots$. Using (4.33) and (4.32) it can be shown that

$$(4.34) \quad \lim_{p \rightarrow \infty} \int_0^{t_0} \sum_{i=1}^m b_i(s, x(s, u_p))u_p^i(s) ds \\ = \int_0^{t_0} \sum_{i=1}^m b_i(s, x(s, u_0))u_0^i(s) ds.$$

It will now be established that

$$\begin{aligned}
 (4.35) \quad & \liminf_{p \rightarrow \infty} \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) u_p^i(s) u_p^j(s) \, ds \\
 & \geq \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_0)) u_0^i(s) u_0^j(s) \, ds.
 \end{aligned}$$

Define $r_p = u_p - u_0$, $p = 1, 2, 3, \dots$, then $r_p \rightarrow 0$ (wk) as $p \rightarrow \infty$, and $u_p = r_p + u_0 \rightarrow u_0$ (wk) as $p \rightarrow \infty$. Then from Theorem 4.3(ii), we have the inequality

$$\begin{aligned}
 (4.36) \quad & \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) u_p^i(s) u_p^j(s) \, ds \\
 & \geq 2 \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) r_p^i(s) u_0^j(s) \, ds \\
 & \quad + \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) u_p^i(s) u_p^j(s) \, ds.
 \end{aligned}$$

It is a simple matter to prove

$$\begin{aligned}
 (4.37) \quad & \lim_{p \rightarrow \infty} 2 \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) r_p^i(s) u_0^j(s) \, ds \\
 & = \lim_{p \rightarrow \infty} 2 \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_0)) r_p^i(s) u_0^j(s) \, ds.
 \end{aligned}$$

Since $r_p \rightarrow 0$ (wk) as $p \rightarrow \infty$, it follows from (4.37) that

$$(4.38) \quad \lim_{p \rightarrow \infty} \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) r_p^i(s) u_0^j(s) \, ds = 0.$$

From (4.38) one may deduce that

$$\begin{aligned}
 (4.39) \quad & \lim_{p \rightarrow \infty} 2 \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) r_p^i(s) u_0^j(s) \, ds \\
 & + \lim_{p \rightarrow \infty} \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) u_0^i(s) u_0^j(s) \, ds \\
 & = \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_0)) u_0^i(s) u_0^j(s) \, ds.
 \end{aligned}$$

By (4.36) and (4.39), we have that

$$\begin{aligned}
 (4.40) \quad & \liminf_{p \rightarrow \infty} \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_p)) u_p^i(s) u_p^j(s) \, ds \\
 & \geq \int_0^{t_0} \sum_{i,j=1}^m c_{ij}(s, x(s, u_0)) u_0^i(s) u_0^j(s) \, ds.
 \end{aligned}$$

It can be shown that

$$(4.41) \quad \lim_{p \rightarrow \infty} \int_0^{t_0} k(s, x(s, u_p), u_p(s)) ds = \lim_{p \rightarrow \infty} K(u_p, t_p).$$

Therefore by (4.31), (4.34), (4.40) and (4.41), we obtain that

$$\gamma = \liminf K(u_p, t_p) \geq K(u_0, t_0) \geq \gamma$$

with $(u_0, t_0) \in \hat{U}(\Omega, \Gamma)$. Thus the proof for Case 1 is complete.

We omit the details for Case 2.

It should be pointed out that the optimal control problems we have considered in this section are of such a general nature that Pontryagin's maximal principle [22] does not apply to give necessary conditions for a minimum. It is, however, interesting to note that, for example, in the fixed time, fixed endpoint special case of Theorem 4.2 ($\Omega =$ constant compact convex set) with the function k of the form $\phi_1(t, x) + \phi_2(t, u)$ (ϕ_1 differentiable in x) then Mangasarian [20, Corollary 2, p. 149] has shown that Pontryagin's maximal principle is both a necessary and a sufficient condition for an optimal controller.

In the case when k is continuous and of the form $k(t, x, u) = f(t, x) + h(t, u)$ with f convex in x , and h convex in u for each t and $f(t, x) \geq 0$, $h(t, u) \geq a |u|^p$ for some $a > 0$, $p > 0$, Lee and Markus [16, Chap. 3] have established an optimal controller. In this case they also give necessary and sufficient conditions for an optimal control.

5. Application. Let $(R^n, \|\cdot\|)$ be the usual Euclidean space of n -dimensions, where for $x = (x^1, x^2, \dots, x^n) \in R^n$, we define $\|x\|^2 = \sum (x^i)^2$. For $\epsilon > 0$, define J_ϵ to be the set

$$J_\epsilon = \{(x, y) \mid x, y \in R^n, \|x - y\| < \epsilon\}.$$

The collection $\mathcal{J} = \{J_\epsilon \mid \epsilon > 0\}$ is the uniformity on R^n induced by the norm $\|\cdot\|$. We define

$$(5.1) \quad \lambda(A, B) = \inf \{ \|x - y\| \mid x \in A, y \in B \},$$

where A and B are subsets of R^n .

Let $[a, b] \subset R$ be a compact interval. Suppose F is a mapping, $F: [a, b] \rightarrow C(R^n)$. If $F(t)$ is nonempty when $a \leq t \leq b$, then we define a mapping $k: [a, b] \times R^n \rightarrow R$ by the equation

$$(5.2) \quad k(t, x) = \lambda(\{x\}, F(t)),$$

where λ is defined by (5.1).

We have the following theorem.

THEOREM 5.1. *If F is usc on $[a, b]$, then $k: [a, b] \times R^n \rightarrow R$ is lsc on $[a, b] \times R^n$.*

Proof. It is well known that for fixed t in $[a, b]$ the function $k(t, \cdot): R^n \rightarrow R$ is continuous [11, p. 100]. It will now be demonstrated that $k(\cdot, x)$ is lsc on $[a, b]$ (uniformly with respect to $x \in R^n$). Since F is usc on $[a, b]$, then for t_0 in $[a, b]$ and $\epsilon > 0$, there is a $\delta_{\epsilon, t_0} > 0$, such that

$$(5.3) \quad t \in [a, b], |t - t_0| < \delta_{\epsilon, t_0} \text{ imply } F(t) \subset J_{\epsilon/2}[F(t_0)].$$

Thus if t is in $[a, b]$, and $|t - t_0| < \delta_{\epsilon, t_0}$, then there is a $y_{t, x}$ in $F(t)$ such that $k(t, x) + \epsilon/2 > \|x - y_{t, x}\|$, and by (5.3) there is a b_{t_0} in $F(t_0)$ such that $\|y_{t, x} - b_{t_0}\| < \epsilon/2$. By the definition of $k(t_0, x)$, we have

$$k(t_0, x) \leq \|x - b_{t_0}\| \leq \|x - y_{t, x}\| + \|y_{t, x} - b_{t_0}\|.$$

Thus if t is in $[a, b]$, and $|t - t_0| < \delta_{\epsilon, t_0}$, then

$$k(t_0, x) < k(t, x) + \epsilon/2 + \epsilon/2 = k(t, x) + \epsilon,$$

whence $k(\cdot, x)$ is lsc at t_0 , and the above $\delta_{\epsilon, t_0} > 0$ depends only on ϵ and t_0 . Finally, it must be established that $k: [a, b] \times R^n \rightarrow R$ is lsc. Thus, given $t_0 \in [a, b]$ and $\epsilon > 0$, pick $\delta_{\epsilon, t_0} > 0$ such that

$$(5.4) \quad t \in [a, b], |t - t_0| < \delta_{\epsilon, t_0} \text{ imply } k(t, x) > k(t_0, x) - \epsilon/2$$

for x in R^n . Let x_0 be an element of R^n , then because $k(t_0, \cdot)$ is continuous at x_0 , there is a $\delta_{t_0, x_0, \epsilon}^* > 0$ such that

$$(5.5) \quad \|x - x_0\| < \delta_{t_0, x_0, \epsilon}^* \text{ implies } |k(t_0, x_0) - k(t_0, x)| < \epsilon/2.$$

Define $\delta_1 = \min(\delta_{\epsilon, t_0}, \delta_{t_0, x_0, \epsilon}^*)$; then $\|x - x_0\| < \delta_1$ and $|t - t_0| < \delta_1$, t in $[a, b]$, imply

$$(5.6a) \quad k(t, x) > k(t_0, x) - \epsilon/2,$$

$$(5.6b) \quad k(t_0, x) > k(t_0, x_0) - \epsilon/2.$$

Combining (5.6a) and (5.6b), we obtain that k is lsc at $(t_0, x_0) \in [a, b] \times R^n$.

THEOREM 5.2. *If F is lsc on $[a, b]$, then k is usc on $[a, b] \times R^n$.*

The proof is similar to that given for Theorem 5.1.

It is interesting to note that "partial" converses of Theorems 5.1 and 5.2 are also true, as is shown by the following.

THEOREM 5.3. *If $k(\cdot, x): [a, b] \rightarrow R$ is lsc (uniformly with respect to x in R^n), then F is usc on $[a, b]$.*

Proof. Suppose t_0 is in $[a, b]$, and $\epsilon > 0$ is given. Then there is a $\delta_{\epsilon, t_0} > 0$ such that

$$t \in [a, b], |t - t_0| < \delta_{\epsilon, t_0} \text{ imply } k(t, x) > k(t_0, x) - \epsilon$$

for $x \in R^n$. Thus if t is an element of $[a, b]$ and $|t - t_0| < \delta_{\epsilon, t_0}$, and x is in $F(t)$, then $k(t, x) = \inf \{\|x - y\| \mid y \in F(t)\} = 0$. Whence $0 \leq k(t_0, x)$

$< \epsilon$. Consequently, there is a $y \in F(t_0)$ such that $\|x - y\| < \epsilon$ (by the definition of $k(t_0, x)$). Therefore, $x \in J_\epsilon[F(t_0)]$ and we have thereby proved

$$t \in [a, b], |t - t_0| < \delta_{\epsilon, t_0} \quad \text{imply} \quad F(t) \subset J_\epsilon[F(t_0)].$$

This completes the proof.

We also obtain the following theorems.

THEOREM 5.4. *If $k(\cdot, x): [a, b] \rightarrow R$ is usc on $[a, b]$ (uniformly with respect to $x \in R^n$), then F is lsci.*

The proof is similar to that given for Theorem 5.3.

THEOREM 5.5. *If $F(t)$ is convex for each t in $[a, b]$, then k is a convex function in the variable x in R^n .*

Proof. See [3, p. 5].

We are now able to generalize a problem stated by Pontryagin, Boltyanskii, Gamkrelidze and Mischenko [22, p. 197 ff.] concerning the application of the mathematical theory of optimal control to a problem in the approximation of functions. In this connection consider the linear optimal control problem with the "moving target set" $F: [0, T] \rightarrow C(R^n)$ with the properties: F is usc on $[0, T]$, $F(t)$ is compact, convex, and nonempty for each t in $[0, T]$. Then by (H_4) of §4,

$$\mathfrak{B} = \{x \mid x = x(t, u), (u, t_1) \in U(\Omega, \Gamma), 0 \leq t \leq t_1\}$$

is a bounded subset of R^n . It may be demonstrated that the mapping k defined in (5.2), $k: [0, T] \times \text{co } \mathfrak{B} \rightarrow R$ (where $\text{co } \mathfrak{B}$ denotes the convex hull of the set \mathfrak{B}) satisfies (H_1) . Instead of using F to define $\hat{U}(\Omega, \Gamma)$ as was done in (3.5), we use $F^*: [0, T] \rightarrow C(R^n) :: t \rightarrow F^*(t) = R^n$ to define $\hat{U}(\Omega, \Gamma)$ in (3.5). If (H_8) is satisfied, then clearly $\hat{U}(\Omega, \Gamma)$ is nonempty. If $h: [0, T] \times R^n \times R^m \rightarrow R$ is a mapping satisfying (H_1) and (H_6') , then $k + h$ is a mapping of the same type satisfying (H_1) and (H_6') . We therefore define $K: U(\Omega, \Gamma) \rightarrow R$ by

$$(5.7) \quad K(u, t_1) = \int_0^{t_1} k(s, x(s, u)) + h(s, x(s, u), u(s)) \, ds$$

for $(u, t_1) \in U(\Omega, \Gamma)$. Theorem 4.2 may then be applied to give that K has an absolute minimum on $\hat{U}(\Omega, \Gamma)$. It appears that $h(t, x, u) = \|u\|^2$ or $h \equiv 0$ would be reasonable choices for cost functionals. Both Theorems 4.2 and 4.3 may be applied to these cases. The case $h \equiv 0$ is especially interesting because in this case the possibility of actually synthesizing an optimal controller is quite promising. The reader will no doubt perceive several other variations of this cost functional to which our results may be applied. The problem mentioned above [22, p. 197 ff.], which Pontryagin and his associates considered, was the special case when $F(t) = \{y(t)\} = \text{singleton point set}$, where, for example, $y: [0, T] \rightarrow R$ is required to be

continuous and the cost functional is

$$\int_0^T [x(t) - y(t)]^2 dt.$$

This functional is to be minimized on the class of all functions $x: [0, T] \rightarrow R$ which have continuous derivatives up to and including those of order n , and the n th derivative $x^{(n)}$ satisfies a Lipschitz condition with constant α .

Acknowledgment. The author wishes to express his appreciation to Professors G. M. Ewing and W. T. Reid for their encouragement and guidance.

REFERENCES

- [1] P. ALEXANDROFF AND H. HOPF, *Topologie*, vol. I, Springer-Verlag, Berlin, 1935.
- [2] A. V. BALAKRISHNAN, *An operator theoretic formulation of a class of control problems and a steepest descent method of solution*, this Journal, 1 (1963), pp. 109–127.
- [3] P. T. BATEMAN, H. RADSTROM, O. HANNER, A. M. MACBEATH, C. A. ROGERS AND V. L. KLEE, *Seminar on convex sets*, The Institute for Advanced Study, Princeton, 1949–1950.
- [4] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.
- [5] L. CESARI, *An existence theorem in problems of optimal control*, this Journal, 3 (1965), pp. 7–22.
- [6] ———, *Existence theorems for optimal solutions in Pontryagin and Lagrange problems*, this Journal, 3 (1965), pp. 475–498.
- [7] A. CHANG, *An optimal regulator problem*, this Journal, 2 (1965), pp. 220–233.
- [8] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [9] A. F. FILIPPOV, *On Certain Questions in the Theory of Optimal Control*, this Journal, 1 (1963), pp. 76–84.
- [10] F. HAUSDORFF, *Set Theory*, 2nd ed., Chelsea, New York, 1962.
- [11] S. T. HU, *Elements of General Topology*, Holden-Day, San Francisco, 1964.
- [12] A. N. KOLMOGOROV AND V. J. FOMIN, *Elements of Functional Analysis*, vol. 1, Graylock Press, Albany, New York, 1957.
- [13] ———, *Elements of Functional Analysis*, vol. 2, Graylock Press, Albany, New York, 1961.
- [14] J. P. LASALLE, *The time optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. V, Princeton University Press, Princeton, 1960, pp. 1–24.
- [15] E. B. LEE AND L. MARKUS, *Optimal control for nonlinear processes*, Arch. Rational Mech. Anal., 8 (1961), pp. 36–58.
- [16] ———, *Foundations of Optimal Control*, to be published.
- [17] L. A. LIUSTERNIK AND V. J. SOBELEV, *Elements of Functional Analysis*, Ungar, New York, 1961.
- [18] E. J. MCSHANE, *Integration*, Princeton University Press, Princeton, 1944.
- [19] E. J. MCSHANE AND T. A. BOTTS, *Real Analysis*, Van Nostrand, Princeton, 1959.
- [20] O. L. MANGASARIAN, *Sufficient conditions for the optimal control of nonlinear systems*, this Journal, 4 (1966), pp. 139–152.

- [21] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, J. Math. Anal. Appl., 7 (1963), pp. 110-117.
- [22] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [23] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Ungar, New York, 1955.
- [24] E. ROXIN, *The existence of optimal controls*, Michigan Math. J., 9 (1962), pp. 109-119.

RELAXED CONTROLS AND VARIATIONAL PROBLEMS*

E. J. McSHANE†

Introduction. In the study of the problems of minimizing or maximizing a functional of a curve on some assigned class of curves, the symbolism of control theory has shown itself to be valuable both in the theory and in the applications. One great advantage of the formulation is that it greatly facilitates the investigation of minimizing curves in which the controls are at some times set at the boundary of the region to which they are confined. All the classical single-integral problems of the calculus of variations can be easily rephrased as optimal control problems, with at worst a mild restriction on the side-equations in Lagrange and Bolza problems [1], [2]. However, most of the theory of optimal control, though not all (cf. [5], [21]), has been developed under the hypothesis that the controls are bounded. Certainly this is enough for a great number of useful applications. Nevertheless, even for applications it would be better to dispense with this restriction; and when simple and familiar problems of the classical calculus of variations are reformulated as optimal control problems, it often fails to hold.

When for each point in state-space the attainable time-derivatives of the state form a bounded closed convex set, an optimal path exists, as has been known for a long time [17], [8]. When the convexity requirement fails we can replace the attainable set by its convex hull, producing the “relaxed control” problem. But it is far from easy to find the necessary conditions satisfied by optimal paths in this case, especially if controls are unbounded. (Warga did this for bounded controls [19], [20].)

Another device became available in 1937, when Young devised objects called “generalized curves”; these include ordinary curves, and also other objects which may be regarded as a correctly-defined substitute for the unrealizable idea of a curve whose derivative is changing incessantly and with infinite rapidity from one to another of a given set of values. These have proved to be of great utility in proving theorems on the existence of solutions of problems with possibly nonlinear side differential equations. Young showed in [22]–[24] the existence of solutions of problems in which the curves are generalized and there are no side conditions. This has been extended to problems with side conditions, both in traditional form [11] and in optimal-control form [6].

However, in order to characterize the solution and to find properties that

* Received by the editors June 27, 1966, and in revised form December 22, 1966.

† Department of Mathematics, University of Virginia, Charlottesville, Virginia. This research was carried out while the author was Principal Investigator under Army Research Office Grant ARO-D-31-124-G662.

would enable us to construct it, we need to find conditions satisfied by all minimizing curves. Furthermore, these conditions should be established without any analogue of the hypothesis that the minimizing curve is "normal", as in earlier treatments of the calculus of variations. For if our theorems need that hypothesis, and we find that the minimum exists and that there is just one "normal" curve satisfying the necessary conditions, we have not removed the possibility that quite a different curve is minimizing but is not normal and so does not satisfy the necessary conditions. Such necessary conditions, including the classical Euler-Lagrange, Weierstrass and Legendre conditions, were established for minimizing generalized curves in 1940 [12] but only with a restriction that can be expressed as "the controls used are interior to the set of usable controls". For problems in optimal-control form, they have been established by Warga [19], [20] who does not make the restriction of interiority. His theorem also applies to problems in which the permitted trajectories are restricted to lie in some subset B of space, and the minimizing trajectory meets the boundary of B ; such problems do not come under the present Theorem 4.7 in Part II of this paper.

Since we have not solved the originally stated problem of finding an (ordinary) curve that minimizes the given functional, but instead have solved the somewhat similar problem of finding a minimizing generalized curve, it behooves us to show that we have, at least in some sense and under auspicious circumstances, solved the original problem. This we did in 1940 [13], and do again in Part III, by showing that under not too restrictive hypotheses on the data of the problem, the minimum is furnished by a generalized curve that happens to be ordinary. Warga establishes a different connection between the problems; knowledge of the minimizing generalized curve permits the construction of nearby ordinary curves for which the functional is arbitrarily close to its minimum.

At the time of preparing the manuscript of this paper the publications on optimal controls had been confined to the case of bounded controls; our results are not restricted in this way. But before the manuscript was submitted, two papers appeared which treated the unbounded case, one by Warga [21] and one by Cesari [3]; I had overlooked these, and owe thanks to the referee for pointing them out. As a result, part of §13 has been heavily revised. However, there is a danger that the unavoidable complications of the unbounded controls may obscure what I consider the more important aspect of the theorems, namely, that without highly sophisticated mathematics existence theorems are proved that are not restricted to linear or convex controls and cost-functions. We presuppose familiarity with the Lebesgue integral for functions of a single variable on a finite interval, but avoid multiple integrals, Lebesgue-Stieltjes integrals and all but very elementary general topology. For the sake of readers who wish to avoid the

extra complications of the unbounded-control case, we have enclosed many statements in heavy brackets. These are needed if the control set is permitted to be unbounded; if it is bounded, the proofs can be shortened by omitting all bracketed statements.

I. EXISTENCE THEOREM

1. Statement of the problem. The standard problem of optimal control can be expressed in a variety of ways. Temporarily ignoring questions of integrability, etc., we can formulate it thus. We are given a set U , the "control set," a subset B of n -dimensional space R^n , and a set of functions $f^1(x, t, u), \dots, f^n(x, t, u)$ defined for all x in B , all real t and all u in U . A function $u(t): t_0 \leq t \leq t_1$ is a "control function"; a "state function" or "trajectory" $x(t) = (x^1(t), \dots, x^n(t))$, $t_0 \leq t \leq t_1$, corresponds to $u(t)$ if $x(t)$ is in B for all t and

$$(1.1) \quad x^i(t) = x^i(t_0) + \int_{t_0}^t f^i(x(\tau), \tau, u(\tau)) d\tau.$$

We are also given an "end set" E in R^{2n+2} and a function $e(x_0, t_0, x_1, t_1)$ on E (here t_0, t_1 are real numbers and x_0, x_1 are n -tuples). The problem is to find a control function $u_0(t)$, $t_0 \leq t \leq t_1$, and a state function $x_0(t)$ corresponding to $u_0(t)$ having $(x_0(t_0), t_0, x_0(t_1), t_1)$ in E and giving to e its least value on all such pairs $(u(t), x(t))$.

If u_1, \dots, u_k are points of U , and during a short time interval $[t, t + \delta]$ the controls are set on u_1 for time $w_1\delta, \dots$, on u_k for time $w_k\delta$, $w_1, \dots, w_k \geq 0$, $\sum w_j = 1$, then the integral of $f^j(x(t), t, u(t))$ from t to $t + \delta$ is nearly

$$(1.2) \quad \int_t^{t+\delta} \sum w_i f^j(x(t), t, u_i) dt.$$

We idealize this concept of rapidly-varying controls by introducing the concept of "relaxed controls". We replace the weighted average $\sum w_i f^j(x(t), t, u_i)$ by any mean-value process $\mathfrak{M}[\cdot, t]$, depending in general on t ; and as in (1.2) we use this mean value as the integrand, instead of $f^j(x(\tau), \tau, u(\tau))$. Thus the integral in (1.1) will be replaced by

$$(1.3) \quad \int_{t_0}^t \mathfrak{M}[f^i(x(\tau), \tau, u), \tau] d\tau.$$

The simplest kind of weighted average of a function is that in which all the weight is given to a single point $u(t)$, so that $\mathfrak{M}[\phi(u), t] = \phi(u(t))$ for every ϕ . With this special kind of mean, (1.3) reduces to the integral in (1.1). So (1.3) is a generalization of (1.1), not just an analogue.

We now begin to state our notational conventions and the precise requirements of the problem. Statements enclosed by [] may be omitted if one is willing to assume that u is bounded and closed.

U is a bounded [or unbounded] closed set in r -dimensional space R^r ; its points will be denoted by $u = (u^1, \dots, u^r)$, with or without affixes.

B is a set in R^n ; its points will be denoted by $x = (x^1, \dots, x^n)$, with or without affixes.

E is a closed set of points (x_0, t_0, x_1, t_1) , where x_0 and x_1 are in B and t and t_1 are real numbers.

(1.4) f^1, \dots, f^n are continuous functions $f^i(x, t, u)$, x in B , t real, u in U .

e is defined and continuous on E .

Given any topological space S , we denote by $C_0[S]$ the set of all real-valued functions defined and continuous on S and vanishing outside a compact subset of S . Thus $C_0[U]$ is the set of all continuous functions on U when U is bounded [but if U is unbounded, $C_0[U]$ is the set of continuous functions $\phi(u)$ such that the set $\{u: \phi(u) \neq 0\}$ is in some sphere].

The usual requirement that B be closed is omitted; it is replaced by a very slightly weaker hypothesis in the main theorem, namely, Theorem 2.7.

To any one familiar with probability theory it is enough to say that the mean-value operator \mathfrak{M} is an expected-value operator, that is, an integration with respect to a probability measure. However, a more elementary theory is sufficient for our needs. To begin with, we suppose that \mathfrak{M} assigns a number $\mathfrak{M}[\phi]$ to each ϕ in $C_0[U]$, and that this has the properties

$$(1.5) \quad \mathfrak{M}[a_1\phi_1 + a_2\phi_2] = a_1\mathfrak{M}[\phi_1] + a_2\mathfrak{M}[\phi_2]$$

whenever ϕ_1 and ϕ_2 are in $C_0[U]$ and a_1 and a_2 are real numbers. We suppose also that

$$(1.6) \quad \mathfrak{M}[\phi] \geq 0 \quad \text{if } \phi \text{ is in } C_0[U] \text{ and } \phi \geq 0.$$

We shall often use the technically incorrect symbol $\mathfrak{M}[\phi(u)]$ to denote $\mathfrak{M}[\phi]$.

[For unbounded U we extend this to a larger class of ϕ . We first construct a sequence ρ_1, ρ_2, \dots of continuous functions on U such that

$$(1.7) \quad \begin{aligned} 0 &\leq \rho_q(u) \leq 1, \quad u \text{ in } U, \quad q = 1, 2, 3, \dots, \quad i = 1, \dots, n, \\ \rho_q(u) &= 1 \quad \text{if } |u^i| \leq q, \\ \rho_q(u) &= 0 \quad \text{if } |u^i| \geq q + 1 \end{aligned}$$

for any of the numbers $i = 1, \dots, n$.

These are all in $C_0[U]$. So if ψ is any function continuous on U , both $\psi\rho_q$ and

$|\psi|_{\rho_q}$ are in $C_0[U]$, $k = 1, 2, 3, \dots$. As k increases, $\mathfrak{M}[|\psi|_{\rho_q}]$ is nondecreasing. If this has a finite limit we say that ψ has an \mathfrak{M} -mean value. In that case the limit of $\mathfrak{M}[\psi_{\rho_q}]$ exists, since the sequences $\mathfrak{M}[|\psi| + \psi]_{\rho_q}$ and $\mathfrak{M}[|\psi| - \psi]_{\rho_q}$ are nondecreasing; and we define

$$(1.8) \quad \mathfrak{M}[\psi] = \lim_{q \rightarrow \infty} \mathfrak{M}[\psi_{\rho_q}].$$

It is easy to verify that this limit does not depend on the choice of the ρ_q .

If we wished to assume only that U is topological, we would here assume the existence of a sequence $\rho_1 \leq \rho_2 \leq \dots$ of functions in $C_0[U]$ satisfying the first of equations (1.7) and having $\lim_{q \rightarrow \infty} \rho_q(u) = 1$ for all u in U .

It is quite easy to prove that if ϕ_1 and ϕ_2 have \mathfrak{M} -mean values, and a_1 and a_2 are real numbers, then $a_1\phi_1 + a_2\phi_2$ has an \mathfrak{M} -mean value, and (1.5) holds; also, (1.6) is obviously true for all ϕ having \mathfrak{M} -mean values.]

We now add one requirement, without which \mathfrak{M} could hardly be called a mean-value operator, namely

$$(1.9) \quad \mathfrak{M}[1] = 1.$$

Next we generalize the idea of control function.

A *relaxed control* function is a function which assigns to each t in an interval $[t_0, t_1]$ a mean-value process $\mathfrak{M}[\cdot, t]$ such that if $\phi(t, u)$, $t_0 \leq t \leq t_1$, u in U , is bounded and continuous, the function $\mathfrak{M}[\phi(t, u), t]$, $t_0 \leq t \leq t_1$, is Lebesgue measurable.

Before stating the next definition, we announce the intention of misusing the integral sign. If $f(t)$ is defined for almost all t in an interval $[a, b]$, say for all t in M , where $[a, b] - M$ has measure 0, and if f has an integral over M , we shall understand

$$\int_a^b f(t) dt = \int_M f(t) dt,$$

in spite of the fact that the integrand in the left member is meaningless on a set of measure 0.

A function $x(t) = (x^1(t), \dots, x^n(t))$, $t_0 \leq t \leq t_1$, is said to be a *trajectory* corresponding to a relaxed control $\mathfrak{M}[\cdot, t]$, $t_0 \leq t \leq t_1$, if the points $(x(t), t)$, $t_0 \leq t \leq t_1$, are in B , and each $f^i(x(t), t, u)$ has an \mathfrak{M} -mean value for almost all t in $[t_0, t_1]$, and the equations

$$(1.11) \quad x^i(t) = x^i(t_0) + \int_{t_0}^t \mathfrak{M}[f^i(x(\tau), \tau, u), \tau] d\tau$$

are satisfied. Since we shall always wish to consider relaxed controls and

corresponding trajectories together, we give the name “generalized curve” to triples $C: (\mathfrak{M}, x(\cdot), [t_0, t_1])$, where $[t_0, t_1]$ is an interval of real numbers, $(\mathfrak{M}[\cdot, t]: (t_0 \leq t \leq t_1))$ is a relaxed control function, and $x(t)$, $t_0 \leq t \leq t_1$, is a trajectory corresponding to the relaxed control function.

A generalized curve $(\mathfrak{M}, x(\cdot), [t_0, t_1])$ is called *admissible*
 (1.12) if $(x(t_0), t_0, x(t_1), t_1)$ is a point of E . The class of all admissible generalized curves will be denoted by \mathfrak{F} .

When discussing a generalized curve $C: (\mathfrak{M}, x(\cdot), [t_0, t_1])$ the corresponding point $(x(t_0), t_0, x(t_1), t_1)$ in R^{2n+2} is mentioned so often that it deserves a symbol. We shall call it the *ends* of C , and denote it by $\eta(C)$:

$$(1.13) \quad \eta(C) = (x(t_0), t_0, x(t_1), t_1).$$

We define

$$(1.14) \quad e_{\min} = \infimum \text{ of } e(\eta(C)) \text{ for all } C \text{ in } \mathfrak{F}.$$

This exists, finite or $-\infty$, if \mathfrak{F} is not empty; but it need not be true that $e(\eta(C)) = e_{\min}$ for any C in \mathfrak{F} .

Now the optimal relaxed-control problem is the problem of finding a member $C_0 = (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ of the class of admissible generalized curves which gives to $e(\eta(C))$ the value e_{\min} .

2. Statement of the existence theorem. By definition of e_{\min} there surely exists a sequence C_1, C_2, \dots of admissible generalized curves, where

$$(2.1) \quad C_k = (\mathfrak{M}_k, x_k(\cdot), [t_{0,k}, t_{1,k}]),$$

such that the ends $\eta(C_k)$ satisfy

$$(2.2) \quad \lim_{k \rightarrow \infty} e(\eta(C_k)) = e_{\min}.$$

Every such sequence is called a minimizing sequence, regardless of convergence.

We shall assume that the following holds:

(2.3) There exists a minimizing sequence C_1, C_2, \dots , whose trajectories all lie in a bounded closed subset of B .

In many applications this is easily seen to be the case. For example, the conditions of the problem may guarantee that B is closed, that all curves in \mathfrak{F} have the same initial point and that for every minimizing sequence the trajectories have bounded length. In such cases reference to a general theorem yielding (2.3) as conclusion would be rather silly. However, it is perhaps worth stating the following theorem.

THEOREM 2.4. Let B be closed; let \mathfrak{F} contain a generalized curve C_1 , and let

E_1 be the subset of E on which $e(x_0, t_0, x_1, t_1) \leq e(\eta(C_1))$. Assume that there is an interval $[a, b]$ that contains t_0 and t_1 whenever (x_0, t_0, x_1, t_1) is in E_1 , and that each i in the set $\{1, \dots, n\}$ is of one of the following two types.

Type 1. The numbers x_0^i, x_1^i are bounded for all (x_0, t_0, x_1, t_1) in E_1 , and there exists a summable nonnegative function $g^i(t)$, $a \leq t \leq b$, such that

$$(2.4a) \quad f^i(x, t, u) \geq -g^i(t)[1 + (x^i)^2]^{1/2}$$

for all x in B , all t in $[a, b]$ and all u in U .

Type 2. Either x_0^i or x_1^i is bounded for all (x_0, t_0, x_1, t_1) in E_1 , and there exists a number N_i such that

$$(2.4b) \quad |f^i(x, t, u)| \leq N_i \sum \{f^j(x, t, u) + g^j(t)[1 + (x^j)^2]^{1/2} + 1\}$$

for all x in B , all t in $[a, b]$ and all u in U , the sum being taken over all j of Type 1 and the g^j being the functions in (2.4a).

Then (2.3) is satisfied.

Given any minimizing sequence C_1, C_2, C_3, \dots , any C_k with $e(\eta(C_k)) > e(\eta(C))$ can be replaced by C_1 , and we thus obtain a sequence with $e(\eta(C_k)) \leq e(\eta(C_1))$ for all k . Let C be any member of \mathfrak{F} with $e(\eta(C)) \leq e(\eta(C_1))$. If i is of Type 1, by (1.11) we have

$$dx^i/dt \geq -g^i(t)[1 + x^i(t)^2]^{1/2}.$$

So if we define

$$y_i(t) = \log \{x^i(t) + [1 + x^i(t)^2]^{1/2}\},$$

we find $dy_i/dt \geq -g^i(t)$, and for all t in $[a, b]$,

$$y_i(t) - y_i(t_0) \geq -\int_a^b g^i(t) dt,$$

$$y_i(t_1) - y_i(t) \geq -\int_a^b g^i(t) dt.$$

Hence $y_i(t)$ is bounded, and so therefore is $x^i(t)$. From this and (1.11), the values of the integral

$$\int_{t_0}^t \left\{ \Re[f^i(x(\tau), \tau, u), \tau] + g^i(\tau)[1 + x^i(\tau)^2]^{1/2} \right\} d\tau$$

for all C with $e(\eta(C)) \leq e(\eta(C_1))$ are bounded, say $\leq K_i$.

From this last and (2.4a) we see that for any subinterval t', t'' of $[t_0, t_1]$ we have, for i of Type 2,

$$\begin{aligned}
 |x^i(t'') - x^i(t')| &\leq \int_{t'}^{t''} \mathfrak{M}[|f^i(x(\tau), \tau, u)|, \tau] d\tau \\
 (2.5) \qquad &\leq \int_{t_0}^{t_1} N_i \sum \{ \mathfrak{M}[f^j(x(\tau), \tau, u), \tau] \\
 &\quad + g^j(\tau)[1 + x^j(\tau)^2]^{1/2} + 1 \} d\tau \\
 &\leq N_i \sum [K_j + (b - a)].
 \end{aligned}$$

Since by hypothesis either $x^i(t_0)$ or $x^i(t_1)$ is bounded, $x^i(t)$ is bounded on $[t_0, t_1]$, and the proof is complete.

[As a very simple application, consider the problem in which $U = R^1$, $B = R^2$, $t_0 = x_0^1 = x_0^2 = x_1^1 = 0$, $t_1 = 1$, $f^1(x, t, u) = u$, $f^2(x, t, u) = [1 + (u)^2]^{1/2}$, $e(x_0, t_0, x_1, t) = x_1^2$, which is merely the problem of finding the shortest path joining $(0, 0)$ and $(1, 0)$. The hypotheses of Theorem 2.4 are satisfied, f^2 being of Type 1 (with $g^2 = 0$) and f^1 being of Type 2 (with $N_1 = 1$). However, the hypothesis

$$x^i f^i(x, t, u) \leq g(t)[1 + \sum (x^i)^2]^{1/2}$$

with g summable, which has been used in theorems involving bounded U , is not satisfied in this simple instance.]

We shall occasionally use the solecism " E is independent of t_0 " to mean that the indicator function of E is independent of t_0 ; that is, if (x_0, t_0', x_1, t_1) is in E , so is (x_0, t_0'', x_1, t_1) for all numbers t_0'' . The corresponding expression is also used with reference to the other coordinates $x_0^1, \dots, x_0^n, x_1^1, \dots, x_1^n, t_1$ of points of E . We use a similar expression for B .

[For the case of unbounded U we shall need a device for comparing the rates of growth of functions as the distance of u from the origin tends to ∞ . If $\psi(x, t, u)$ and $G(x, t, u)$ are defined for all (x, t) in a set A and all u in U , we say that ψ is of slower growth than G (or G of faster growth than ψ) uniformly on A , provided that $G \geq 0$ and for each positive ϵ there is a bounded subset U_ϵ of U such that

$$(2.6) \qquad |\psi(x, t, u)| \leq \epsilon G(x, t, u)$$

whenever (x, t) is in A and u is in $U - U_\epsilon$.]

We can now state our existence theorem; the hypotheses in heavy brackets are unnecessary, and in fact are automatically satisfied, whenever U is bounded.

THEOREM 2.7. *With the notation and assumptions listed in (1.4), assume that there exists a minimizing sequence C_1, C_2, \dots , where $C_k = (\mathfrak{M}_k, x_k(\cdot), [t_{0,k}, t_{1,k}])$, whose trajectories all lie in a bounded closed subset A of B . [Assume further that there exists a continuous function $G(x, t, u)$, (x, t) in A , u in U ,*

such that

(i) for almost all t in $[t_{0,k}, t_{1,k}]$ the \mathfrak{M}_k -mean value $\mathfrak{M}_k[G(x_k(t), t, u), t]$ exists, and the integrals

$$\int_{t_{0,k}}^{t_{1,k}} \mathfrak{M}_k[G(x_k(\tau), \tau, u), \tau] d\tau, \quad k = 1, 2, 3, \dots,$$

exist and are bounded;

(ii) G is of faster growth than 1 uniformly on A ;

(iii) for each number j in the set $\{1, \dots, n\}$, either (a) f^j is of slower growth than G uniformly on A , or else (b) $f^j \geq 0$, and f^1, \dots, f^n and B are independent of x^j , and E is independent of x_1^j , and e is a monotonic nondecreasing function of x_1^j for fixed values of the other variables.]

Then the class \mathfrak{F} of admissible generalized curves contains a member $C_0 = (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ such that $e(\eta(C_0)) = e(x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0})$ is equal to the infimum e_{\min} of the values of $e(\eta(C))$ for all curves C in \mathfrak{F} .

3. Proof of the theorem. [There is clearly no loss of generality in assuming, as we shall, that

$$(3.1) \quad G(x, t, u) \geq 1, \quad (x, t) \text{ in } A, \quad u \text{ in } U.]$$

By hypothesis, there is an interval $[a, b]$ that contains all the intervals $[t_{0,k}, t_{1,k}]$, $k = 1, 2, 3, \dots$. To simplify notation we define $\mathfrak{M}_k[\cdot, t]$ to be 0 if $a \leq t < t_{0,k}$ or if $t_{1,k} < t \leq b$; then our integrals can be taken over any subinterval of $[a, b]$, but we must keep in mind that \mathfrak{M}_k is not a relaxed control function except for t in $[t_{0,k}, t_{1,k}]$. Correspondingly, we understand $x_k(t)$ to mean $x_k(t_{0,k})$ if $a \leq t < t_{0,k}$ and to mean $x_k(t_{1,k})$ if $t_{1,k} < t \leq b$.

For the sake of brevity, in the remainder of this proof the letter t will always denote a point of the interval $[a, b]$.

Since the ends of the trajectory of C_k lie in the bounded set A , it is possible to select a convergent subsequence. We shall suppose that C_1, C_2, \dots is already such a subsequence, and define

$$(3.2) \quad t_\alpha = \lim_{k \rightarrow \infty} t_{\alpha,k}, \quad x_\alpha^i = \lim_{k \rightarrow \infty} x_k^i(t_{\alpha,k}), \quad \alpha = 0, 1, \quad i = 1, \dots, n.$$

The averaging operator \mathfrak{M}_0 for the minimizing curve will be defined by operating under the integral sign; we first integrate \mathfrak{M}_k over intervals $[a, t]$, then we pass to the limit (which behaves the way we would like the integral of \mathfrak{M}_0 to behave), then we define \mathfrak{M}_0 by differentiating the integral. In the process, we shall frequently use the following notation. Given any continuous function $\psi(x, t, u)$, we define for each positive integer k ,

$$(3.3) \quad \psi^{[k]}(t) = \int_a^t \mathfrak{M}_k[\psi(x_k(\tau), \tau, u), \tau] d\tau, \quad a \leq t \leq b,$$

provided that the integral exists. Also, we define

$$(3.4) \quad \hat{\psi}(t) = \lim_{k \rightarrow \infty} \psi^{[k]}(t), \quad a \leq t \leq b,$$

provided that the limit exists.

[If ψ is either G or any of the f^i for which (b) of Theorem 2.7 (iii) is satisfied, the corresponding $\psi^{[k]}$ are nondecreasing and all have a common bound. So by Helly's theorem we can select a subsequence that converges at each point t in $[a, b]$. We suppose that C_1, C_2, \dots is already such a subsequence. Then (3.4) can be applied, and

$$(3.5) \quad \hat{G}(t) = \lim_{k \rightarrow \infty} G^{[k]}(t), \quad \hat{f}^i(t) = \lim_{k \rightarrow \infty} f^{i[k]}(t), \quad a \leq t \leq b,$$

for all i such that f^i satisfies (b).]

If we combine (3.5), (3.3), (3.2) and (1.11) we see that

$$(3.6) \quad \begin{aligned} \hat{f}^i(t) &= \lim_{k \rightarrow \infty} \{x_k^i(t) - x_{k,0}^i\} \\ &= \lim_{k \rightarrow \infty} x_k^i(t) - x_0^i \end{aligned}$$

for all i such that f^i satisfies (b).]

If ϕ is in $C_0[A \times U]$, $|\phi|$ has a finite upper bound, which we call B_ϕ . Then

$$(3.7) \quad |\phi^{[k]}(t'') - \phi^{[k]}(t')| \leq \left| \int_{t'}^{t''} \mathfrak{M}_k[B_\phi, \tau] d\tau \right| = B_\phi |t'' - t'|.$$

So all the $\phi^{[k]}$ satisfy the same Lipschitz condition, and by Ascoli's theorem we can select a subsequence of C_1, C_2, \dots for which the $\phi^{[k]}$ are uniformly convergent on $[a, b]$. Moreover, for the limit $\hat{\phi}$ we obtain from (3.7),

$$(3.8) \quad |\hat{\phi}(t'') - \hat{\phi}(t')| \leq B_\phi |t'' - t'|.$$

There exists a countable subset Φ of $C_0[A \times U]$ such that every function ψ in $C_0[A \times U]$ can be uniformly approximated by functions in Φ . By the diagonal process we can select a subsequence of C_1, C_2, \dots , such that for every ϕ in Φ the corresponding sequence of $\phi^{[k]}$ is uniformly convergent. We suppose C_1, C_2, \dots to be already such a subsequence; then (3.4) defines $\hat{\phi}$ for every ϕ in Φ , and moreover the convergence is uniform, and (3.8) holds.

[For unbounded U we also need the following estimate.

If ψ is of slower growth than G uniformly on A , for each positive ϵ there is a ϕ in Φ , such that

$$(3.9) \quad |\psi(x, t, u) - \phi(x, t, u)| < \epsilon G(x, t, u)$$

for all (x, t, u) in $A \times U$.

Define ψ_0 on $A \times U$ by setting

$$\psi_0(x, t, u) = \begin{cases} \psi(x, t, u) & \text{where } |\psi(x, t, u)| \leq (\epsilon/2)G(x, t, u), \\ (\epsilon/2)G(x, t, u) & \text{where } \psi(x, t, u) > (\epsilon/2)G(x, t, u), \\ -(\epsilon/2)G(x, t, u) & \text{where } \psi(x, t, u) < -(\epsilon/2)G(x, t, u). \end{cases}$$

Then $\psi - \psi_0$ is continuous and is 0 except on a bounded set. This implies that there is a ϕ in Φ such that

$$|(\psi - \psi_0) - \phi| < \epsilon/2 \quad \text{on } A \times U.$$

Hence

$$\begin{aligned} |\psi - \phi| &\leq |\psi_0| + |\psi - \psi_0 - \phi| \\ &< (\epsilon/2)G(x, t, u) + \epsilon/2 \\ &< \epsilon G(x, t, u), \end{aligned}$$

and (3.9) is established.]

For the case of bounded U , the next proof is to be read with the simplification that $G(x, t, u)$ is to be replaced by 1.

(3.10) Let ψ be continuous on $A \times U$ [and of slower growth than G uniformly on A]. Then the functions $\psi^{[k]}$ converge uniformly on $[a, b]$.

The numbers $2(b - a) + 2G^{[k]}(b)$ have by hypothesis a finite upper bound; let K be such a bound.

Let ϵ be positive. Since Φ is dense in $C_0[A \times U]$ [and (3.9) is valid], there is a ϕ in Φ such that

$$|\psi(x, t, u) - \phi(x, t, u)| < \epsilon K^{-1}G(x, t, u)$$

for all (x, t) in A and u in U . Since $\phi^{[k]}$ converges uniformly to $\hat{\phi}$ on $[a, b]$, there is a k_ϵ such that if $k > k_\epsilon$,

$$|\phi^{[k]}(t) - \hat{\phi}(t)| < \epsilon/K, \quad a \leq t \leq b.$$

Then if $h, k > k_\epsilon$ and $a \leq t \leq b$,

$$\begin{aligned} &|\psi^{[h]}(t) - \psi^{[k]}(t)| \\ &\leq |\psi^{[h]}(t) - \phi^{[h]}(t)| + |\phi^{[h]}(t) - \phi^{[k]}(t)| + |\phi^{[k]}(t) - \psi^{[k]}(t)| \\ &< 2 \int_a^b \epsilon K^{-1} \mathfrak{N}_k[G(x_k(\tau), \tau, u), \tau] d\tau + 2\epsilon K^{-1}(b - a) \\ &< \epsilon, \end{aligned}$$

which implies the uniform convergence of the $\psi^{[k]}$.

(3.11) If ψ is continuous on $A \times U$ [and is of slower growth than G uniformly on A], $\hat{\psi}$ is absolutely continuous.

For bounded U we have already established in (3.8) the stronger conclusion that $\hat{\psi}$ is Lipschitzian. [Let ϵ be positive and ϕ as in (3.9) but with $\epsilon[1 + 2\hat{G}(b)]^{-1}$ in place of ϵ in (3.9). Denote by B_ϕ a positive upper bound for $|\phi|$. Define $\delta = \epsilon/(2B_\phi)$. Let $[a_1, b_1), \dots, [a_h, b_h)$ be disjoint subintervals of $[a, b]$ with total length less than δ . Define $\gamma = \psi - \phi$. Then for $k = 1, 2, 3, \dots$, we have, by (3.9) and (3.7),

$$\begin{aligned} & \sum_{j=1}^h |\psi^{[k]}(b_j) - \psi^{[k]}(a_j)| \\ & \leq \sum_{j=1}^h \{|\phi^{[k]}(b_j) - \phi^{[k]}(a_j)| + |\gamma^{[k]}(b_j) - \gamma^{[k]}(a_j)|\} \\ & \leq B_\phi \sum_{j=1}^h [b_j - a_j] + \epsilon(1 + 2\hat{G}(b))^{-1} \int_a^b \mathfrak{M}_k[G(x_k(\tau), \tau, u), \tau] d\tau. \end{aligned}$$

Letting $k \rightarrow \infty$ yields

$$\sum_{j=1}^h |\hat{\psi}(b_j) - \hat{\psi}(a_j)| < B_\phi(\epsilon/(2B_\phi)) + \epsilon/2 = \epsilon,$$

proving the absolute continuity of $\hat{\psi}$.]

For each i [such that f^i satisfies hypothesis (a) of Theorem 2.7 (iii)], we can apply (3.11) and (3.10); the functions $f^i(t)$, $a \leq t \leq b$, are defined and absolutely continuous, and the $f^{i[k]}$ converge uniformly to them. With the x_0^i of (3.2) we now define

$$(3.12) \quad y^i(t) = x_0^i + f^i(t), \quad i = 1, \dots, n.$$

[These need not be the x_0^i of the minimizing generalized curve; later we shall modify them if necessary.] As in (3.6) we see that now

$$(3.13) \quad y^i(t) = \lim_{k \rightarrow \infty} x_k^i(t), \quad i = 1, \dots, n, \quad a \leq t \leq b,$$

by (3.10) the convergence is uniform [if f^i satisfies (a) of Theorem 2.7 (iii)]. This, with (3.2), implies that $(y^i(t), t)$, $a \leq t \leq b$, is the limit of the sequence $(x_k^i(t), t)$ of points of A , and that $(y(a), t_0, y(b), t_1)$ is the limit of the sequence $(x_k(a), t_{0,k}, x_k(b), t_{1,k})$ of points of E . Since A and E are closed, these limits belong to A and E respectively. Since e is continuous,

$$(3.14) \quad e(y(a), t_0, y(b), t_1) = \lim_{k \rightarrow \infty} e(x_k(t_{0,k}), t_{0,k}, x_k(t_{1,k}), t_{1,k}) = e_{\min}.$$

Each of the functions \hat{G} , x_0^j and $\hat{\phi}$ (ϕ in Φ) is of bounded variation, hence has a derivative on all of $[a, b]$ except a set of measure 0. Since Φ is counta-

ble, the union of these sets has measure 0. On the rest of $[a, b]$, which we call M , all the functions G, x_0^j and $\hat{\phi}, \phi$ in Φ , have finite derivatives. We now extend this.

(3.15) If ψ is continuous on $A \times U$ [and is of slower growth than G uniformly on A] and t_2 is in M , then $\psi'(t_2)$ exists.

(For bounded U the following proof can be simplified by replacing G by 1 and $G^{[k]}(t)$ and $\hat{G}(t)$ by t .)

Let ϵ be positive. Let ϕ be in Φ and satisfy the inequality in (3.9); define $\gamma = \psi - \phi$. Suppose $a < t_2 < b$; the modifications for $t_2 = a$ or $t_2 = b$ are obvious. Then

$$\begin{aligned} |\gamma^{[k]}(t) - \gamma^{[k]}(t_2)| &\leq \epsilon \int_{t_2}^t \mathfrak{N}_k[G(x_k(\tau), \tau, u), \tau] d\tau \\ &= \epsilon |G^{[k]}(t) - G^{[k]}(t_2)|. \end{aligned}$$

Letting $k \rightarrow \infty$ yields

$$|\hat{\gamma}(t) - \hat{\gamma}(t_2)| \leq \epsilon |\hat{G}(t) - \hat{G}(t_2)|.$$

If $|t - t_2|$ is small enough and not equal to 0,

$$|[\hat{\phi}(t) - \hat{\phi}(t_2)]/(t - t_2) - \hat{\phi}'(t_2)| < \epsilon$$

and

$$|[\hat{G}(t) - \hat{G}(t_2)]/(t - t_2) - \hat{G}'(t_2)| < 1,$$

whence

$$|[\hat{\psi}(t) - \hat{\psi}(t_2)]/(t - t_2) - \hat{\phi}'(t_2)| < \epsilon(|\hat{G}'(t_2)| + 2).$$

So the values of $[\hat{\psi}(t) - \hat{\psi}(t_2)]/(t - t_2)$ for small positive $|t - t_2|$ differ from each other by an arbitrarily small amount, and the derivative $\hat{\psi}'(t_2)$ exists.

(3.16) If ψ is continuous [and of bounded support] on $A \times U$ and t_2 is in M , and $\psi(y(t_2), t_2, u) \geq 0$ for all u in U , then $\hat{\psi}'(t_2) \geq 0$.

Since ψ is in $C_0[A \times U]$, it is uniformly continuous, so to each $\epsilon > 0$ corresponds $\delta > 0$, such that if (x, t) is in A and $|t - t_2| < \delta$ and $|x^i - y^i(t_2)| < \delta, i = 1, \dots, n$, then

$$(3.17) \quad \psi(x, t, u) > \psi(y(t_2), t_2, u) - \epsilon \geq -\epsilon$$

for all u in U . We now wish to show that

(3.18) there is a positive γ less than δ and there is a positive integer k_0 such that if $|t - t_2| < \gamma$ and $k > k_0$, then

$$|x_k^i(t) - y^i(t_2)| < \delta.$$

[We distinguish two cases. Suppose that f^i satisfies (a) of Theorem 2.7 (iii).] By (3.11), $y^i(\cdot)$ is continuous, so that

$$(3.19) \quad \text{there is a positive } \gamma \text{ such that} \\ |y^i(t) - y^i(t_2)| < \delta/2 \quad \text{if} \quad |t - t_2| < \gamma.$$

By (3.13) and the sentence following it, there is a k_0 such that if $k > k_0$,

$$|y^i(t) - x_k^i(t)| < \delta/2, \quad a \leq t \leq b.$$

These two inequalities establish (3.18). [Suppose next that f^i satisfies (b) of Theorem 2.7 (iii). Since $y^{i'}(t_2)$ exists, y^i is continuous at t_2 , and (3.19) holds. Choose t', t'' such that $t_2 - \gamma < t' < t_2 < t'' < t_2 + \gamma$. By (3.13), there is a k_0 such that if $k > k_0$, then

$$(3.20) \quad |y^i(t') - x_k^i(t')| < \delta/2 \quad \text{and} \quad |y^i(t'') - x_k^i(t'')| < \delta/2.$$

Since x_k^i is nondecreasing by (1.11) and (b) of Theorem 2.7 (iii), by (3.19) and (3.20) we have

$$y^i(t_2) - \delta < x_k^i(t') \leq x_k^i(t) \leq x_k^i(t'') < y^i(t_2) + \delta$$

for all t in (t', t'') and all $k > k_0$. So (3.18) holds with γ replaced by the smaller of $t'' - t_2$ and $t_2 - t'$.]

Now by (3.17) and (3.18), if $t_2 < t < t_2 + \gamma$,

$$\begin{aligned} \psi^{[k]}(t) - \psi^{[k]}(t_2) &= \int_{t_2}^t \mathfrak{M}_k[\psi(x_k(\tau), \tau, u), \tau] d\tau \\ &\geq \int_{t_2}^t (-\epsilon) d\tau \\ &= -\epsilon(t - t_2). \end{aligned}$$

Letting $k \rightarrow \infty$ yields

$$\hat{\psi}(t) - \hat{\psi}(t_2) \geq -\epsilon(t - t_2).$$

Dividing by $t - t_2$ and letting $t \rightarrow t_2$ yields

$$\hat{\psi}'(t_2) \geq -\epsilon.$$

Since ϵ is arbitrary, the proof is complete.

We can now define the mean-value operator \mathfrak{M}_0 for the minimizing generalized curve C_0 by means of the following lemma.

LEMMA 3.21. *For each ϕ in $C_0[U]$ and each t_2 in M , the derivative $\hat{\psi}'(t_2)$ has one and the same value for all ψ in $C_0[A \times U]$ such that $\psi(y(t_2), t_2, u) = \phi(u)$, u in U . This common value is denoted by $\mathfrak{M}_0[\phi, t_2]$.*

If ψ_1 and ψ_2 are in $C_0[A \times U]$ and

$$\psi_i(y(t_2), t_2, u) = \phi(u), \quad i = 1, 2, \quad u \text{ in } U,$$

then $\psi_1 - \psi_2$ and $\psi_2 - \psi_1$ both satisfy the hypotheses of (3.16), so both $\hat{\psi}_1 - \hat{\psi}_2$ and $\hat{\psi}_2 - \hat{\psi}_1$ have nonnegative derivatives at t_2 ; that is, $\hat{\psi}_1'(t_2) = \hat{\psi}_2'(t_2)$, which proves the assertion in Lemma 3.21.

From Lemma 3.21 we immediately deduce that if ψ is in $C_0[A \times U]$, then for all t_2 in M ,

$$(3.22) \quad \mathfrak{M}_0[\psi(y(t_2), t_2, u), t_2] = \hat{\psi}'(t_2).$$

[Furthermore,

let ψ be continuous on $A \times U$. Then (a) if ψ is of slower growth than G uniformly on A , then $\psi(y(t_2), t_2, u)$ has an $\mathfrak{M}_0[\cdot, t_2]$ -mean value, and (3.22) holds for all t_2 in M , and (b) if $\psi \geq 0$, and $\hat{\psi}(b)$ exists and is finite, then $\psi(y(t), t, u)$ has an $\mathfrak{M}_0[\cdot, t]$ -mean value for almost all t in M , and

$$\int_a^b \mathfrak{M}_0[\psi(y(t), t, u), t] dt \leq \hat{\psi}(b).$$

For each positive integer q let ρ_q be as in (1.7). Define $\phi_q = \psi \cdot \rho_q$, $\psi_q = \psi - \phi_q$. By (2.6), if $\epsilon > 0$ we have

$$-\epsilon G < \psi_q < \epsilon G$$

for all large q . It follows that $\epsilon \hat{G} + \hat{\psi}_q$ and $\epsilon \hat{G} - \hat{\psi}_q$ are nondecreasing. Hence

$$(3.24) \quad -\epsilon \hat{G}'(t_2) \leq \hat{\psi}_q'(t_2) \leq \epsilon \hat{G}'(t_2).$$

From $\psi = \psi_q + \phi_q$, we deduce $\hat{\psi} = \hat{\psi}_q + \hat{\phi}_q$, whence

$$\hat{\psi}'(t_2) = \hat{\psi}_q'(t_2) + \hat{\phi}_q'(t_2).$$

By (3.24), the last term tends to 0 as $q \rightarrow \infty$, so by (3.22),

$$(3.25) \quad \begin{aligned} \hat{\psi}'(t_2) &= \lim_{q \rightarrow \infty} \hat{\psi}_q'(t_2) \\ &= \lim_{q \rightarrow \infty} \mathfrak{M}_0[\psi(y(t_2), t_2, u) \rho_q(u), t_2]. \end{aligned}$$

The same argument applied to $|\psi|$ shows that $\mathfrak{M}_0[|\psi| \rho_q, t_2]$ has a finite limit as $q \rightarrow \infty$. Hence ψ has an \mathfrak{M}_0 -mean value, which by definition is the last expression in (3.25); so (3.22) is satisfied and (a) in (3.23) is proved.

For (b) in (3.23) we observe that for all q and k ,

$$\psi_q^{[k]}(b) \leq \psi^{[k]}(b),$$

whence

$$(3.26) \quad \hat{\psi}_q(b) \leq \hat{\psi}(b).$$

By (3.22),

$$(3.27) \quad \int_a^b \mathfrak{N}_0[\psi(y(t), t, u)\rho_q(u), t] dt \leq \hat{\psi}(b)$$

for all q . By the monotone convergence theorem,

$$\lim_{q \rightarrow \infty} \mathfrak{N}_0[\psi(y(t), t, u)\rho_q(u), t] < \infty$$

for almost all t . For such t , by definition this limit is $\mathfrak{N}_0[\psi(y(t), t, u), t]$. Again by the monotone convergence theorem, with (3.26) and (3.27),

$$\begin{aligned} \hat{\psi}(b) &\geq \lim_{q \rightarrow \infty} \int_a^b \mathfrak{N}_0[\psi(y(t), t, u)\rho_q(u), t] dt \\ &= \int_a^b \mathfrak{N}_0[\psi(y(t), t, u), t] dt, \end{aligned}$$

establishing (b) in (3.23).]

\mathfrak{N}_0 is still undefined on $[a, b] - M$; to complete the definition we choose some t_3 in $M \cap (t_0, t_1)$ and define $\mathfrak{N}_0[\cdot, t] = \mathfrak{N}_0[\cdot, t_3]$ whenever t is in $[a, b] - M$. In the special case $\psi = 1$, we compute

$$\psi^{[k]}(t) = \begin{cases} 0 & \text{if } a \leq t \leq t_{0,k}, \\ t - t_{0,k} & \text{if } t_{0,k} < t \leq t_{1,k}, \\ t_{1,k} - t_{0,k} & \text{if } t_{1,k} < t \leq b. \end{cases}$$

Hence

$$\hat{\psi}(t) = \begin{cases} 0 & \text{if } a \leq t \leq t_0, \\ t - t_0 & \text{if } t_0 < t \leq t_1, \\ t_1 - t_0 & \text{if } t_1 < t \leq b, \end{cases}$$

and so

$$\mathfrak{N}_0[1, t] = \begin{cases} 0 & \text{if } a \leq t < t_0 \text{ or } t_1 < t \leq b, \\ 1 & \text{if } a < t < b, \text{ } t \text{ in } M. \end{cases}$$

By our completion of \mathfrak{N}_0 , the last holds for all t in $[t_0, t_1]$.

Now \mathfrak{N}_0 has properties (1.5) and (1.6), by Lemma 3.21, and property (1.9) if $t_0 \leq t \leq t_1$. Also, if $(\phi(t, u): t_0 \leq t \leq t_1, u \text{ in } U)$ is bounded and continuous, $\mathfrak{N}_0[\phi(t, u), t]$ is measurable by Lemma 3.21 and (3.22), since derivatives are measurable functions. So \mathfrak{N}_0 is a relaxed control function on $[t_0, t_1]$.

We now define

$$(3.28) \quad x_0^i(t) = x_0^i + \int_{t_0}^t \mathfrak{M}_0[f^i(y(\tau), \tau, u), \tau] d\tau, \quad a \leq t \leq b;$$

the integral exists by (3.22) and (3.11) [and (3.23)]. Also [for each j such that f^j satisfies (a) of Theorem 2.7 (iii)], by (3.11) and (3.22), with (3.3), (3.2) and (3.13),

$$\begin{aligned} x_0^j(t) &= x_0^j + \hat{f}^j(t) - \hat{f}^j(t_0) \\ &= \lim_{k \rightarrow \infty} \left\{ x_k^j(t_{0,k}) + \int_{t_{0,k}}^t \mathfrak{M}_k[f^j(x_k(\tau), \tau, u), \tau] d\tau \right\} \\ (3.29) \quad &= \lim_{k \rightarrow \infty} x_k^j(t) \\ &= y^j(t), \end{aligned} \quad a \leq t \leq b.$$

[If f^j satisfies (b) of Theorem 2.7 (ii), B and the $f^i(x, t, u)$, $i = 1, \dots, n$, are all independent of x^j , and substituting $x_0^j(\tau)$ for $y^j(\tau)$ in (3.28) has no effect.] So for $i = 1, \dots, n$, the function x^i satisfies

$$(3.30) \quad x_0^i(t) = x_0^i + \int_{t_0}^t \mathfrak{M}_0[f^i(x_0(\tau), \tau, u), \tau] d\tau, \quad a \leq t \leq b.$$

We saw in the sentences preceding (3.14) that $(y^i(t), t)$, $a \leq t \leq b$, is in A (hence in B) and $(y(a), t_0, y(b), t_1)$ is in E . But by (3.29), $y^j(a) = x_0^j(a) = x_0^j(t_0)$ for all j , and $y^j(b) = x_0^j(b) = x_0^j(t_1)$ for all j [for which (a) of Theorem 2.7 (iii) holds. For the other j , E and e are independent of x_1^j], so $(x_0(t_0), t_0, x_0(t_1), t_1)$ is in E , and by (3.14),

$$\begin{aligned} e(x_0(t_0), t_0, x_0(t_1), t_1) &= e(y(a), t_0, y(b), t_1) \\ &= e_{\min}. \end{aligned}$$

Likewise, if $t_0 \leq t \leq t_1$, $(y^j(t), t)$ is in B , and $y^j(t) = x_0^j(t)$ for all j [for which (a) of Theorem 2.7 (iii) holds. For the others, B is independent of x^j], so $(x_0(t), t)$ is in B . Therefore $C_0 = (\mathfrak{M}_0, x_0(\cdot), [t_0, t_1])$ is an admissible generalized curve; and $e(\eta(C_0)) = e_{\min}$; and the theorem is established.

II. NECESSARY CONDITIONS FOR AN OPTIMUM

4. Statement of the theorem. Although no differentiability was required of the f^i in establishing the existence of an optimizing generalized curve in Part I, such conditions are at least highly desirable in establishing necessary conditions. As before, statements in heavy brackets may be omitted if one is willing to assume that U is bounded and closed.

We retain the notation of Part I, and throughout §§4–8 make the following two assumptions.

(4.1) $C_0 = (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ is a member of the class \mathfrak{F} of admissible generalized curves such that

$$e(x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0}) = e_{\min}.$$

(4.2) There is a positive number A such that whenever u is in U and (x, t) is in R^{n+1} and there exists a t' in $[t_{0,0}, t_{1,0}]$ for which $|t - t'| \leq A$ and $\|x - x_0(t')\| \leq A$, the point x is in B , and the derivatives $f_t^i, f_{x_j}^i, i, j = 1, \dots, n$, are defined and continuous at (x, t, u) . [Also, there exists a nonnegative continuous function $G(t, u), t_{0,0} \leq t \leq t_{1,0}, u$ in U , such that $\mathfrak{M}_0[G(t, u), t]$ is finite for almost all t and is summable over $[t_{0,0}, t_{1,0}]$, and for $i, j = 1, \dots, n$ and (x, t, u) satisfying the first part of the assumption, the inequalities

$$|f_{x_j}^i(x, t, u)| \leq G(t', u), \quad |f_t^i(x, t, u)| \leq G(t', u)$$

are satisfied.]

The "variations" used in the discussion involve change of mean-value operators and shifting of the endpoint $\eta(C_0) = (x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0})$. We consider the latter first. A vector $(\xi_1, \tau_1, \xi_1, \tau_1)$ (in which ξ_0 and ξ_1 each have n components) can reasonably be said to be directed "into" E at $\eta(C_0)$ if there is a curve $(X_0(z), T_0(z), X_1(z), T_1(z)), 0 \leq z \leq \epsilon$, beginning at $\eta(C_0)$, lying in E and having tangent vector $(\xi_0, \tau_0, \xi_1, \tau_1)$ at $z = 0$. However, we need the ability to compute with several such vectors simultaneously, so we ask more. We shall use the symbol E' to denote a collection of $(2n + 2)$ -component vectors with the following property.

(4.3) If $(\xi_{0,\sigma}, \tau_{0,\sigma}, \xi_{1,\sigma}, \tau_{1,\sigma}), \sigma = 1, \dots, s$, is any finite set of vectors belonging to E' , there are $2n + 2$ functions $X_0^i(z), T_0(z), X_1^i(z), T_1(z), i = 1, \dots, n, z = (z_1, \dots, z_s)$, defined and continuously differentiable on a neighborhood of $z = 0$ with the properties:

- (a) if all z_σ are ≥ 0 , $(X_0(z), T_0(z), X_1(z), T_1(z))$ is in E ;
- (b) $(X_0(0), T_0(0), X_1(0), T_1(0)) = (x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0})$;
- (c) at $z = 0$ the following equations are satisfied:

$$\partial X^i / \partial z_\sigma = \xi_{\alpha,\sigma}^i, \quad \partial T_\alpha / \partial z_\sigma = \tau_{\alpha,\sigma},$$

$$\alpha = 0, 1, \quad i = 1, \dots, n, \quad \sigma = 1, \dots, s.$$

Since the above symbol for points in E' is cumbersome, we shall use H as an abbreviation for $(\xi_0, \tau_0, \xi_1, \tau_1)$; if these last have a subscript (as in (4.3)), we attach the same subscript to H . Without loss of generality we

can add the hypothesis:

$$(4.4) \quad \begin{array}{l} E' \text{ is a convex cone; if } H' \text{ and } H'' \text{ are in } E' \text{ and } a' \text{ and } a'' \\ \text{are } \geq 0, \text{ then } a'H' + a''H'' \text{ is in } E'. \end{array}$$

A point u_0 of U is in the *support* of $\mathfrak{N}_0[\cdot, t_2]$ for a given t_2 in $[t_{0,0}, t_{1,0}]$ if, whenever ϕ is continuous and nonnegative, and $\phi(u_0) > 0$, we must have $\mathfrak{N}_0[\phi(u), t_2] > 0$ if it exists. For example, in the simple case in which there is a point u_0 of U such that $\mathfrak{N}_0[\phi(u), t_2] = \phi(u_0)$, we see that this point u_0 is the only one in the support of $\mathfrak{N}_0[\cdot, t_2]$.

The letter λ will be used to designate a point $(\lambda_1, \lambda_2, \dots, \lambda_n)$ of R^n . We now define two functions important in the theory. For each (x, t, u) in the domain of the function f^i and each λ in R^n we define

$$(4.5) \quad F(x, t, u, \lambda) = \sum_{i=1}^n \lambda_i f^i(x, t, u),$$

$$(4.6) \quad M(x, t, \lambda) = \inf_{u \in U} F(x, t, u, \lambda).$$

The latter may be $-\infty$ or finite.

All the results of §§4–8 are contained in the following theorem.

THEOREM 4.7 *Let $C_0 = (\mathfrak{N}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ satisfy conditions (4.1) and (4.2), and let E' satisfy (4.3) and (4.4). Then there exist functions $\lambda_0(t), \dots, \lambda_n(t), t_{0,0} \leq t \leq t_{1,0}$, with which the following statements hold:*

- (i) $\lambda_0(t)$ is either constantly 0 or constantly 1;
- (ii) the $\lambda_i(t)$ do not all vanish at any point t of $[t_{0,0}, t_{1,0}]$;
- (iii) the $\lambda_i(t)$ satisfy the equations

$$\lambda_i(t) = \lambda_i(t_{0,0}) - \int_{t_{0,0}}^t \mathfrak{N}_0[F_{x^i}(x_0(\tau), \tau, u, \lambda(\tau)), \tau] d\tau,$$

$$i = 1, \dots, n, \quad t_{0,0} \leq t \leq t_{1,0};$$

- (iv) there is a (finite) constant c such that the equation

$$M(x_0(t), t, \lambda(t)) = c + \int_{t_{0,0}}^t \mathfrak{N}_0[F_t(x_0(\tau), \tau, u, \lambda(\tau)), \tau] dt$$

holds for all t in $[t_{0,0}, t_{1,0}]$ [except those in a set of measure 0; for all t , the left member is at least equal to the right member];

- (v) for almost all t in $[t_{0,0}, t_{1,0}]$ the equation

$$F(x_0(t), t, u_0, \lambda(t)) = M(x_0(t), t, \lambda(t))$$

holds for all u_0 in the support of $\mathfrak{N}_0[\cdot, t]$;

- (vi) if we denote the right member of the equation in (iv) above by $\mu(t)$,

then for all vectors $(\xi_0, \tau_0, \xi_1, \tau_1)$ in E' the inequality

$$\sum_{j=1}^n \{[(\lambda_0)(\partial e/\partial x_0^j) + \lambda_j(t_{0,0})]\xi_0^j + [(\lambda_0)(\partial e/\partial x_1^j) - \lambda_j(t_{1,0})]\xi_1^j\} \\ + [(\lambda_0)(\partial e/\partial t_0) - \mu(t_{0,0})]\tau_0 + [(\lambda_0)(\partial e/\partial t_1) + \mu(t_{1,0})]\tau_1 \geq 0$$

is satisfied, the partial derivatives being evaluated at $(x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0})$.

5. Variations. We first establish a part of the conclusion under the additional hypothesis:

$$(5.1) \quad \text{For all points } (x_0, t_0, x_1, t_1) \text{ of } E \text{ it is true that } t_0 = t_{0,0} \\ \text{and } t_1 = t_{1,0}.$$

Thus all admissible generalized curves will have the same initial and final values of t that C_0 has, so we can simplify notation slightly by changing $t_{0,0}$ to t_0 and $t_{1,0}$ to t_1 . From (4.3) we see that if (5.1) holds,

$$(5.2) \quad \tau_0 = \tau_1 = 0 \quad \text{for all } (\xi_0, \tau_0, \xi_1, \tau_1) \text{ in } E'.$$

If $\mathfrak{M}_1, \dots, \mathfrak{M}_s$ are relaxed control functions, and c_1, \dots, c_s are non-negative numbers, then for each set z_1, \dots, z_s of nonnegative numbers such that $c_1 z_1 + \dots + c_s z_s \leq 1$, the function

$$(5.3) \quad \mathfrak{M}_z = (1 - c_1 z_1 - \dots - c_s z_s)\mathfrak{M}_0 + c_1 z_1 \mathfrak{M}_1 + \dots + c_s z_s \mathfrak{M}_s$$

is readily seen to be also a relaxed control function.

Now by a "variation" we shall mean a triple

$$v = (c, \mathfrak{M}, H)$$

in which c is a nonnegative number, \mathfrak{M} is a relaxed control with bounded closed support, and H is a vector belonging to E' . Suppose now that v_1, \dots, v_s are variations,

$$(5.4) \quad v_\sigma = (c_\sigma, \mathfrak{M}_\sigma, H_\sigma), \quad \sigma = 1, \dots, s.$$

By (4.3),

$$(5.5) \quad \text{there exist functions } X_0^i(z), T_0(z), X_1^i(z), T_1(z) \text{ defined} \\ \text{and continuously differentiable for all } z_1, \dots, z_s \text{ near } 0, \\ \text{having } (X_0^i(z), T_0(z), X_1^i(z), T_1(z)) \text{ in } E \text{ for all small} \\ \text{nonnegative } z_1, \dots, z_s \text{ (so that, by (5.1), } T_0 \text{ and } T_1 \text{ are} \\ \text{constants), and satisfying at } z = 0 \text{ the equation}$$

$$\partial X_\alpha^i / \partial z_\sigma = \xi_{\alpha, \sigma}^i, \quad \partial T_\alpha / \partial z_\sigma = 0,$$

$$\alpha = 1, 2, \quad \sigma = 1, \dots, s, \quad i = 1, \dots, n.$$

These may not be unique; but if several such X_0 , etc., exist, we choose one such set for each set (v_1, \dots, v_s) and retain it throughout this proof.

We shall now attempt to solve the differential equations

$$(5.6) \quad x^i(t, z) = X_0^i(z) + \int_{t_0}^t \mathfrak{M}_z[f^i(x(\tau, z), \tau, u), \tau] dt$$

for fixed z near $(0, \dots, 0)$, \mathfrak{M}_z being defined by (5.3) for all z . (It will not necessarily be a relaxed control unless the z_σ are small and nonnegative; but it is meaningful for all z and all continuous ϕ that have an \mathfrak{M}_0 -mean value.) For this we have to consider the function

$$(5.7) \quad \mathfrak{M}_z[f^i(x, t, u), t]$$

for t in $[t_0, t_1]$, x near $x_0(t)$, z in R^s . As a function of z for fixed x, t this is linear by (5.3), and for all z ,

$$(5.8) \quad \frac{\partial}{\partial z_\sigma} \mathfrak{M}_z[f^i(x, t, u), t] = c_\sigma \{-\mathfrak{M}_0[f^i(x, t, u), t] + \mathfrak{M}_\sigma[f^i(x, t, u), t]\}.$$

With the A of (4.2), let N be the set of all points (x, t) such that for some t' in $[t_0, t_1]$ it is true that

$$(t - t') \leq A/2, \quad |x^i - x_0^i(t')| \leq A/2, \quad i = 1, \dots, n.$$

If ϕ is in $C_0[N \times U]$ and its derivatives ϕ_{x^i} , $i = 1, \dots, n$, are also in $C_0[N \times U]$, and (x, t) is in N , and $x(h) = (x^1 + h, x^2, \dots, x^n)$, then for $|h|$ small but not 0 we have

$$(5.8a) \quad h^{-1}[\phi(x(h), t, u) - \phi(x(0), t, u)] = \phi_{x^1}(\bar{x}, t, u),$$

where \bar{x} is between $x(0)$ and $x(h)$. The right member vanishes identically outside a bounded closed subset of U , and on that bounded closed set it converges to $\phi_{x^1}(x, t, u)$ uniformly as $h \rightarrow 0$, so

$$(5.9) \quad \lim_{h \rightarrow 0} (h^{-1} \{ \mathfrak{M}_z[\phi(x(h), t, u), t] - \mathfrak{M}_z[\phi(x(0), t, u), t] \}) = \mathfrak{M}_z[\phi_{x^1}(x(0), t, u), t].$$

That is, $\mathfrak{M}_z[\phi(x, t, u), t]$ has a partial derivative with respect to x^1 , and it is given by the right member of (5.9). A like result holds for the other x^j , so

$$(5.10) \quad \frac{\partial}{\partial x^j} \mathfrak{M}_z[\phi(x, t, u), t] = \mathfrak{M}_z[\phi_{x^j}(x, t, u), t], \quad j = 1, \dots, n.$$

For bounded U this implies at once

$$(5.11) \quad \frac{\partial}{\partial x^j} \mathfrak{M}_z[f^i(x, t, u), t] = \mathfrak{M}_z[f_{x^j}^i(x, t, u), t]$$

for (x, t) in N .

[For unbounded U we need to investigate further. Equation (5.11) holds with \mathfrak{M}_σ in place of \mathfrak{M}_z , $\sigma = 1, \dots, s$, because \mathfrak{M}_σ has bounded closed support, and we can replace f by a function in $C_0[N \times U]$ that has continuous partial derivatives with respect to the x^j and coincides with f^i on a neighborhood of $N \times (\text{support of } \mathfrak{M}_\sigma)$. We shall show that (5.11) holds also with \mathfrak{M}_0 in place of \mathfrak{M}_z ; then by (5.3), (5.11) will hold for \mathfrak{M}_z .

Let (x, t) be a point of N such that $G(t, u)$ has an \mathfrak{M}_0 -mean value. For each positive integer q we choose a function ρ_q as in (1.7). Then if $\epsilon > 0$, for all large q we have

$$(5.12) \quad \mathfrak{M}_0[G(t, u)(1 - \rho_q(u)), t] < \epsilon/4.$$

Define

$$\phi = \rho_q \cdot f^i, \quad \psi = (1 - \rho_q) \cdot f^i.$$

With the notation of the first part of this proof, (5.8a) holds, and so does its analogue for ψ . From the latter, together with (5.12) and the estimate

$$|\psi_{x^1}(\bar{x}, t, u) - \psi_{x^1}(x, t, u)| = |\{f_{x^1}^i(\bar{x}, t, u) - f_{x^1}^i(x, t, u)\}(1 - \rho_q(u))| \\ \leq 2G(t, u)(1 - \rho_q(u)),$$

we deduce

$$(5.13) \quad |h^{-1}\{\mathfrak{M}_0[\psi(x(h), t, u), t] - \mathfrak{M}_0[\psi(x(0), t, u), t]\} \\ - \mathfrak{M}_0[\psi_{x^1}(x(0), t, u)]| < \epsilon/2.$$

For all small positive $|h|$, by (5.9) we have

$$|h^{-1}\{\mathfrak{M}_0[\phi(x(h), t, u), t] - \mathfrak{M}_0[\phi(x(0), t, u), t]\} \\ - \mathfrak{M}_0[\phi_{x^1}(x(0), t, u), t]| < \epsilon/2.$$

Since $f^i = \phi + \psi$, this and (5.13) combine to yield

$$|h^{-1}\{\mathfrak{M}_0[f^i(x(h), t, u), t] - \mathfrak{M}_0[f^i(x, t, u), t]\} - \mathfrak{M}_0[f_{x^1}^i(x, t, u), t]| < \epsilon.$$

So (5.11) holds for $j = 1$. By a like proof, it holds for $j = 2, \dots, n$ also.]

Now, by a known theorem [14, p. 356] equations (5.6) have a unique solution $x^i(t, z)$, $t_0 = t \leq t_1$, $c = 1, \dots, n$, for all z near 0, and this solution satisfies

$$(5.14) \quad x^i(t, 0) = x_0^i(t), \quad t_0 \leq t \leq t_1.$$

Also, $x^i(t, z)$ has a partial derivative with respect to each z_σ ; to find the differential equations satisfied by these derivatives we need only differentiate both members of (5.6). It is convenient to introduce a notation for these partial derivatives. Given a variation $v = (c, \mathfrak{M}, H)$, we define

$\xi^i(t | v), i = 1, \dots, n, t_0 \leq t \leq t_1$, as the (unique) solution of the equations

$$(5.15) \quad \begin{aligned} \xi^i(t | v) = & \xi_0^i + \int_{t_0}^t \left\{ -c\mathfrak{M}_0[f^i(x_0(\tau), \tau, u), \tau] \right. \\ & + c\mathfrak{M}[f^i(x_0(\tau), \tau, u), \tau] \\ & \left. + \sum_{j=1}^n \mathfrak{M}_0[f_{x^j}^i(x_0(\tau), \tau, u), \tau] \xi^j(\tau | v) \right\} d\tau. \end{aligned}$$

By formal differentiation of (5.6), the partial derivative $\partial x^i / \partial z_\sigma$ satisfies this equation with $v = v_\sigma$, at $z = 0$. Since the solution is unique,

$$(5.16) \quad \begin{aligned} \partial x^i(t, z) / \partial z_\sigma = \xi^i(t | v_\sigma) \quad \text{at} \quad z = (0, \dots, 0), \\ i = 1, \dots, n, \quad \sigma = 1, \dots, s. \end{aligned}$$

The differential equation (5.15) is important enough in this proof to justify simplification and further study. We define

$$(5.17) \quad \Gamma_j^i(t) = \mathfrak{M}_0[f_{x^j}^i(x_0(t), t, u), t].$$

Then (5.15) is an instance (with proper choice of functions g^i) of the equation

$$(5.18) \quad \xi^i(t) = \xi^i(t_0) + \int_{t_0}^t \left\{ \sum_{j=1}^n \Gamma_j^i(\tau) \xi^j(\tau) + g^i(\tau) \right\} d\tau.$$

For each j in the set $1, \dots, n$, let $\Phi_j^1, \dots, \Phi_j^n$ be a solution of the homogeneous equations

$$(5.19) \quad \phi^i(t) = \phi^i(t_0) + \int_{t_0}^t \sum_{j=1}^n \Gamma_j^i(\tau) \phi^j(\tau) d\tau;$$

we may choose these so that the matrix $(\Phi_j^i(t); i, j = 1, \dots, n)$ is non-singular for $t = t_0$, and then it will remain nonsingular for all t in $[t_0, t_1]$. Let $\Psi(t)$ be its inverse matrix, so that for all t in $[t_0, t_1]$,

$$(5.20) \quad \sum_{j=1}^n \Psi_j^i(t) \Phi_k^j(t) = \sum_{j=1}^k \Phi_j^i(t) \Psi_k^j(t) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases}$$

Then the (unique) solution of (5.18) is given by

$$(5.21) \quad \xi^i(t) = \sum_{j,k=1}^n \Phi_j^i(t) \left\{ \Psi_k^j(t_0) \xi^k(t_0) + \int_{t_0}^t \Psi_k^j(\tau) g^k(\tau) d\tau \right\}.$$

The proof is trivial; for if $\xi^i(t)$ is defined by (5.21), it has value $\xi^i(t_0)$ when $t = t_0$, by (5.20), and its derivative is almost everywhere equal to

$$\begin{aligned}
& \sum_{j,k=1}^n \left[(d\Phi_j^i/dt) \left\{ \Psi_k^j(t_0) \xi^k(t_0) + \int_{t_0}^t \Psi_k^j(\tau) g^k(\tau) d\tau \right\} + \Phi_j^i(t) \Psi_k^j(t) g^k(t) \right] \\
&= \sum_{j,k,h=1}^n [\Gamma_h^i(t) \Phi_j^h(t)] \left\{ \Psi_k^j(t_0) \xi^k(t_0) + \int_{t_0}^t \Psi_k^j(\tau) g^k(\tau) d\tau \right\} + g^i(t) \\
&= \sum_{k,h=1}^n \Gamma_h^i(t) \xi^h(t) + g^i(t),
\end{aligned}$$

so that it satisfies (5.18).

In particular, the solution $\xi(t|v)$ of (5.15) is

$$\begin{aligned}
(5.22) \quad \xi^i(t|v) &= \sum_{j,k=1}^n \Phi_j^i(t) \left\{ \Psi_k^j(t_0) \xi_0^k \right. \\
&\quad \left. + c \int_{t_0}^t \Psi_k^j(\tau) (\mathfrak{M}[f^k(x_0(\tau), \tau, u), \tau] \right. \\
&\quad \left. - \mathfrak{M}_0[f^k(x_0(\tau), \tau, u), \tau]) d\tau \right\}.
\end{aligned}$$

6. A fundamental lemma. In the following definition we use the notation

$$(6.1) \quad v = (c, \mathfrak{M}, H), \quad H = (\xi_0, 0, \xi_1, 0).$$

DEFINITION 6.2. For each variation v , let $Y(v) = (Y^0(v), \dots, Y^n(v))$ be the vector with components

$$\begin{aligned}
Y^0(v) &= \sum_{j=1}^n [(\partial e / \partial x_0^j) \xi_0^j + (\partial e / \partial x_1^j) \xi_1^j], \\
Y^i(v) &= \xi^i(t_1|v) - \xi_1^i,
\end{aligned}$$

where the partial derivatives of e are evaluated at the ends $\eta(C_0)$ of C_0 , and $\xi^i(t|v)$ is the solution of (5.15). The set K is defined to be the set of all $Y(v)$, for all variations v .

LEMMA 6.3. K is a convex cone with vertex at $(0, \dots, 0)$; that is, if $Y(v_1)$ and $Y(v_2)$ are in K , and a_1 and a_2 are nonnegative, there is a variation v such that

$$(6.4) \quad Y(v) = a_1 Y(v_1) + a_2 Y(v_2).$$

Let $v_\sigma = (c_\sigma, \mathfrak{M}_\sigma, H_\sigma)$, $\sigma = 1, 2$. Define $\mathfrak{M} = [a_1 c_1 + a_2 c_2]^{-1} (a_1 c_1 \mathfrak{M}_1 + a_2 c_2 \mathfrak{M}_2)$ unless $a_1 c_1 = a_2 c_2 = 0$, in which case define $\mathfrak{M} = \mathfrak{M}_1$. Define also

$$c = a_1 c_1 + a_2 c_2, \quad H = a_1 H_1 + a_2 H_2, \quad v = (c, \mathfrak{M}, H).$$

Then \mathfrak{M} is a relaxed control, and by (4.4), H is in L' . The solutions $\xi^i(t|v_\sigma)$

of (5.15) are given, according to (5.22), by

$$(6.5) \quad \xi^i(t | v_\sigma) = \sum_{j,k=1}^n \Phi_j^i(t) \left\{ \Psi_k^j(t_0) \xi_{0,\sigma}^k + c_\sigma \int_{t_0}^t \Psi_k^j(\tau) (\mathfrak{M}_\sigma[f^i(x_0(\tau), \tau, u), \tau] - \mathfrak{M}_0[f^i(x_0(\tau), \tau, u), \tau]) d\tau \right\},$$

where $\sigma = 1, 2$. If we multiply by a_σ and sum over $\sigma = 1, 2$ in the right member we obtain the same expression as (6.5) with the subscript σ deleted, which by (5.22), is $\xi^i(t | v)$. Hence $\xi^i(t | v) = a_1 \xi^i(t | v_1) + a_2 \xi^i(t | v_2)$, from which we readily deduce (6.4).

The next lemma is the mainspring of the present proof.

LEMMA 6.6. *The point $(-1, 0, \dots, 0)$ is not interior to K .*

Suppose this false. There then exists a positive number δ such that all points of R^{n+1} with distance at most $\delta\sqrt{n}$ from $(-1, 0, \dots, 0)$ belong to K . In particular, the $n + 1$ points

$$(6.7) \quad (-1, -\delta, 0, \dots, 0), (-1, 0, -\delta, 0, \dots, 0), \dots, (-1, 0, \dots, 0, -\delta), (-1, \delta, \delta, \dots, \delta)$$

are all in K . By definition of K there exist variations v_1, v_2, \dots, v_{n+1} such that the corresponding vectors $Y(v_1), \dots, Y(v_{n+1})$ are the vectors (6.7) in the same order. For all real z we define \mathfrak{M}_z by (5.3), with $s = n + 1$; and we denote by $X_0^i(z), T_0(z), X_1^i(z), T_1(z)$ the functions of (5.5), and by $x(t, z)$ the solutions of (5.6).

Consider now the system of equations

$$(6.8) \quad e(X_0(z), t_0, X_1(z), t_1) + y - e_{\min} = 0, \\ x^i(t_1, z) - X_1^i(z) = 0, \quad i = 1, \dots, n.$$

These have the initial solution

$$(6.9) \quad z_1 = \dots = z_{n+1} = y = 0,$$

by (4.3)(b), (5.14) and (4.1). At this initial solution the partial derivatives of the left members of the $n + 1$ equations (6.8) with respect to z_k are the numbers $Y^0(v_k), \dots, Y^n(v_k)$, by (5.16) and (5.5). These are the elements of a nonsingular matrix, being the numbers (6.7). So by the implicit functions theorem, equations (6.8) have solutions

$$(6.10) \quad z_k = Z_k(y), \quad k = 1, \dots, n + 1,$$

for all real numbers for all y near 0; these are differentiable, and reduce to 0 at $y = 0$.

The equations

$$e(X_0(Z(y)), t_0, X_1(Z(y)), t_1) + y - e_{\min} = 0,$$

$$x^i(t_1, Z(y)) - X_1^i(Z(y)) = 0, \quad i = 1, \dots, n,$$

are identically satisfied near $y = 0$. Differentiating and setting $y = 0$ yields

$$\sum_{k=1}^{n+1} Y^0(v_k) Z_k'(0) + 1 = 0,$$

$$\sum_{k=1}^{n+1} Y^i(v_k) Z_k'(0) = 0, \quad i = 1, \dots, n.$$

But since the $Y(v_1), \dots, Y(v_k)$ are the vectors (6.7), we see by inspection that these have a solution

$$Z_k'(0) = 1/(n+1);$$

and since the matrix of coefficients is nonsingular, this is the only solution. Hence for small positive y , the numbers $Z_k(y)$ are all positive, and are arbitrarily near 0 since $Z_k(0) = 0$. But then for positive y near 0 the z_k defined by (6.10) provide us, via (5.3), with an \mathfrak{M}_z that is a relaxed control. By (5.6), $C_z = (\mathfrak{M}_z, x(\cdot, z), [t_0, t_1])$ is a generalized curve. Since (5.6) and (6.8) hold, we have

$$x^i(t_0, z) = X_0^i(z), \quad x^i(t_1, z) = X_1^i(z),$$

so $\eta(C_z) = (x(t_0, z), t_0, x(t_1, z), t_1)$ is in E , and C_z is in the class \mathfrak{F} of admissible curves; so by definition of e_{\min} we must have $e(x(t_0, z), 0, x(t_1, z), 0) \geq e_{\min}$. But since $y > 0$ this contradicts the first of equations (6.8), and the proof is complete.

COROLLARY 6.11. *There exist numbers $\chi_0 = 0$ or 1, χ_1, \dots, χ_n not all 0, such that $\sum_{i=0}^n \chi_i Y^i(v) \geq 0$ for all vectors $Y(v)$ in K .*

If K has no interior points, it lies in a hyperplane $\sum_{i=0}^n c_i Y^i = 0$. If $c_0 = 0$ we define $\chi_i = c_i$, $i = 0, \dots, n$; if $c_0 \neq 0$ we define $\chi_i = c_i/c_0$, $i = 0, \dots, n$; in either case the conclusion holds. If K has interior points, no point $(-\epsilon, 0, \dots, 0)$, $\epsilon \geq 0$, is interior to K , since K is a cone and $(-1, 0, \dots, 0)$ would be interior to K , contrary to Lemma 6.6. This implies (cf. [4, pp. 19-20]) that there is a hyperplane $c_i Y^i = 0$ such that $c_i Y^i \geq 0$ if Y is in K and $c_i Y^i \leq 0$ if $Y = (-\epsilon, 0, 0, \dots, 0)$. The last implies $c_0 \geq 0$. If $c_0 = 0$, we define $\chi_i = c_i$, $i = 0, \dots, n$; if $c_0 \neq 0$, we define $\chi_i = c_i/c_0$, $i = 0, \dots, n$; in either case, the conclusion is valid.

7. Proof of theorem under supplementary hypothesis. The inequality in Corollary 6.11 can be put into a more convenient form with the help of the functions

$$(7.1) \quad \lambda_i(t) = \sum_{j,k=1}^n \chi_j \Phi_k^j(t_1) \Psi_i^k(t), \quad i = 1, \dots, n, \quad t_0 \leq t \leq t_1.$$

By (5.20), these satisfy

$$(7.2) \quad \lambda_i(t_i) = \chi_i, \quad i = 1, \dots, n.$$

We also define

$$(7.3) \quad \lambda_0(t) = \lambda_0 = \chi_0, \quad t_0 \leq t \leq t_1.$$

Let v be a variation (c, \mathfrak{M}, H) . If we replace the $Y^i(v)$ in Corollary 6.11 by their definitions in Definition 6.2, and in the result replace $\xi^i(t_i | v)$ by its value as given by (5.22), and then make the substitutions (7.1), (7.2), (7.3) and (4.5), we obtain

$$(7.4) \quad \begin{aligned} & \sum_{i=1}^n \{[(\lambda_0)(\partial e / \partial x_0^i) + \lambda_i(t_0)]\xi_0^i + [(\lambda_0)(\partial e / \partial x_1^i) - \lambda_i(t_1)]\xi_1^i\} \\ & + c \int_{t_0}^{t_1} \{\mathfrak{M}[F(x_0(\tau), \tau, u, \lambda(\tau), \tau) - \mathfrak{M}_0[F(x_0(\tau), \tau, u, \lambda(\tau)), \tau]] d\tau \\ & \geq 0. \end{aligned}$$

As a first result, by choosing \mathfrak{M} to be any relaxed control with compact support, c to be 0, and $H = (\xi_0, 0, \xi_1, 0)$ to be any member of E' (cf. (5.2)), we obtain conclusion (vi) of Theorem 4.7 provided, that hypothesis (5.1) is satisfied.

Suppose there is a t at which $\lambda_0, \lambda_1, \dots, \lambda_n$ all are 0. The coefficients of χ_j in (7.1) are the elements of the product of the nonsingular matrices $\Phi(t_i)$ and $\Psi(t)$, so they form a nonsingular matrix. Hence the only solution of (7.1) at t is $\chi_1 = \dots = \chi_n = 0$. With (7.3), this implies that χ_0, \dots, χ_n are all 0, which is false. So conclusions (i) and (ii) of Theorem 4.7 hold.

From (7.1) and (5.20),

$$(7.5) \quad \sum_{i=1}^n \lambda_i(t) \Phi_h^i(t) = \sum \chi_j \Phi_h^j(t_i), \quad h = 1, \dots, n, \quad t_0 \leq t \leq t_1.$$

The functions Φ_j^i are absolutely continuous, since they satisfy (5.19). Hence the elements of the inverse matrix Ψ_j^i are also absolutely continuous, and by (7.1) so are the λ_i . By differentiating (7.5) we obtain for almost all t ,

$$\sum_{i=1}^n \{(\lambda_i/dt) \Phi_h^i(t) + \lambda_i(t) (d\Phi_h^i(t)/dt)\} = 0.$$

For the last factor we substitute its value from (5.19), then multiply by $\Psi_k^h(t)$ and sum on h ; the result is

$$d\lambda_k/dt + \sum_{i=1}^n \lambda_i(t) \Gamma_k^i(t) = 0.$$

Therefore, the λ_i satisfy Theorem 4.7 (iii).

Now let u_1, u_2, u_3, \dots be a countable set of points dense in U . For each positive integer m and each t in $[t_0, t_1]$ we define

$$\mu_m(t) = \min_{j=1, \dots, m} F(x_0(t), t, u_j, \lambda(t)).$$

Each of the m functions of t named in the right member is continuous, so μ_m is also continuous. If we let A_k denote the set of those points t in $[t_0, t_1]$ for which

$$(7.6) \quad \mu_m(t) = F(x_0(t), t, u_k, \lambda(t)),$$

we see that every point of $[t_0, t_1]$ is in at least one of the sets A_k , and the A_k are closed. We define

$$D_1 = A_1, D_2 = A_2 - A_1, \dots, D_m = A_m - [A_1 \cup \dots \cup A_{m-1}].$$

These are disjoint and measurable; their union is $[t_0, t_1]$, and

$$(7.7) \quad (7.6) \text{ holds on } D_k.$$

We now define a function U_m by

$$U_m(t) = u_k \quad \text{for } t \text{ in } D_k, \quad k = 1, \dots, m.$$

This is bounded and measurable. If we define $\mathfrak{N}_m[\phi, t]$ for all continuous ϕ by

$$(7.8) \quad \mathfrak{N}_m[\phi, t] = \phi(U_m(t));$$

this is a relaxed control with bounded support. Now we apply (7.4) to the variation $(1, \mathfrak{N}_m, 0)$ and obtain

$$(7.9) \quad \int_{t_0}^{t_1} \{\mathfrak{N}_m[F(x_0(\tau), \tau, u, \lambda(\tau)), \tau] - \mathfrak{N}_0[F(x_0(\tau), u, \lambda(\tau)), \tau]\} d\tau \geq 0.$$

But each t in $[t_0, t_1]$ is in some D_k , and for this k we have by (7.8), (7.9), and (7.7),

$$\begin{aligned} \mathfrak{N}_m[F(x_0(t), \lambda(t), u), t] &= F(x_0(t), t, u_k, \lambda(t)) \\ &= \mu_m(t). \end{aligned}$$

Hence (7.9) yields

$$(7.10) \quad \int_{t_0}^{t_1} \mu_m(\tau) d\tau \geq \int_{t_0}^{t_1} \mathfrak{N}_0[F(x_0(\tau), \tau, u, \lambda(t)), \tau] d\tau.$$

As m increases, $\mu_m(t)$ is nondecreasing for each t , and by (4.6),

$$(7.11) \quad \lim_{m \rightarrow \infty} \mu_m(t) = M(x_0(t), t, \lambda(t)).$$

By the monotone convergence theorem, (7.10) implies

$$(7.12) \quad \int_{t_0}^{t_1} M(x_0(\tau), \tau, \lambda(\tau)) \, d\tau \geq \int_{t_0}^{t_1} \mathfrak{M}_0[F(x_0(\tau), \tau, u, \lambda(\tau)), \tau] \, d\tau.$$

But for all (t, u) in $[t_0, t_1] \times U$ we have $F(x_0(t), t, u, \lambda(t)) \geq M(x_0(t), t, \lambda(t))$, so

$$\mathfrak{M}_0[F(x_0(t), t, u, \lambda(t)), t] \geq M(x_0(t), t, \lambda(t)).$$

Hence (7.12) can hold only if

$$(7.13) \quad \mathfrak{M}_0[F(x_0(t), t, u, \lambda(t)), t] = M(x_0(t), t, \lambda(t))$$

for almost all t in $[t_0, t_1]$.

Let t be a point at which (7.13) holds, and let U_{\min} be the subset of U on which the function

$$(7.14) \quad F(x_0(t), t, u, \lambda(t)) - M(x_0(t), t, \lambda(t)), \quad u \text{ in } U,$$

vanishes. This function is continuous on U , and by (4.6) is positive on $U - U_{\min}$. Since the \mathfrak{M}_0 -mean of the function (7.14) is 0, no point of $U - U_{\min}$ can be in the support of \mathfrak{M}_0 . So (7.14) vanishes on the support of \mathfrak{M}_0 , and we have established Theorem 4.7(v).

8. Removal of the supplementary hypothesis. Now we turn to the proof of the theorem without the added hypothesis (5.1). The device used is to construct a new problem in which (5.1) holds, but the values of the function e to be minimized are the same as in the original problem. The sets and functions entering in the new problem will be denoted by affixing an asterisk (superscript or subscript) to the corresponding symbol for the original problem, with two exceptions that will be indicated.

The set U^* consists of those points

$$(8.1) \quad u_* = (u_*^1, \dots, u_*^{r+1}) = (u, w)$$

with u in U and w equal to $\frac{1}{2}$ or to $\frac{3}{2}$. As indicated in (8.1), we use u and w as alternate names for (u_*^1, \dots, u_*^r) and u_*^{r+1} respectively. The set B^* will consist of all points

$$(8.2) \quad x_* = (x_*^1, \dots, x_*^{n+1}) = (x, t)$$

of R^{n+1} such that $x = (x_*^1, \dots, x_*^n)$ is in B and t (that is, x_*^{n+1}) is real. The symbol t_* , with or without other affixes, will denote a real number.

The set E_* in R^{2n+4} will consist of all points $(x_{*0}, t_{*0}, x_{*1}, t_{*1})$ with (x_{*0}, x_{*1}) in E and

$$(8.3) \quad t_{*0} = t_{0,0}, \quad t_{*1} = t_{1,0}.$$

The set E_*' will consist of all vectors $(\xi_{*0}, 0, \xi_{*1}, 0)$ with (ξ_{*0}, ξ_{*1})

$= (\xi_0, \tau_0, \xi_1, \tau_1)$ in E' . This evidently satisfies (4.4). To show that it satisfies (4.3), let $(\xi_{*0,\sigma}, 0, \xi_{*1,\sigma}, 0)$, $\sigma = 1, \dots, s$, belong to E_*' . Then the vectors $(\xi_{0,\sigma}, \tau_{0,\sigma}, \xi_{1,\sigma}, \tau_{1,\sigma})$ are in E' ; with the corresponding functions $X_{0,\sigma}, T_{0,\sigma}, X_{1,\sigma}, T_{1,\sigma}$ of (4.3), we define

$$\begin{aligned} X_{*\alpha,\sigma}(z) &= (X_{\alpha,\sigma}(z), T_{\alpha,\sigma}(z)), & \alpha &= 1, 2, \\ T_{*0,\sigma}(z) &= T_{*1,\sigma}(z) = 0. \end{aligned}$$

Thus the functions $(X_{*0,\sigma}, T_{*0,\sigma}, X_{*1,\sigma}, T_{*1,\sigma})$ satisfy the requirements (4.3) for the new problem.

For each point of E_* we define

$$\begin{aligned} (8.4) \quad e_*(x_{*0}, t_{*0}, x_{*1}, t_{*1}) &= e(x_{*0}, x_{*1}) \\ &= e(x_0, t_0, x_1, t_1). \end{aligned}$$

The analogues of f^1, \dots, f^n will be denoted by g^1, \dots, g^{n+1} (the asterisk notation would be typographically inconvenient); they are defined by

$$\begin{aligned} (8.5) \quad g^i(x_*, t_*, u_*) &= g^i(x, t, t_*, u, w) \\ &= wf^i(x, t, u), \quad i = 1, \dots, n+1, \end{aligned}$$

in the last of which we understand that

$$(8.6) \quad f^{n+1}(x, t, u) = 1.$$

(These are independent of t_* .)

The definitions of relaxed control, trajectory, generalized curve and admissible generalized curve are unchanged except in notation. The class of admissible generalized curves will be called \mathfrak{F}^* ; a generalized curve $C^* = (\mathfrak{N}^*, x_*(\cdot), [t_*', t_*''])$ is in \mathfrak{F}^* if and only if $(x_*(t_*'), t_*', x_*(t_*''), t_*'')$ is in E_* (which requires in particular $t_*' = t_{*0} = t_{0,0}$ and $t_*'' = t_{*1} = t_{0,0}$), and the equations

$$(8.7) \quad x_*^i(t_*) = x_*^i(t_{*0}) + \int_{t_{*0}}^{t_*} \mathfrak{N}^*[g^i(x_*(\sigma), \sigma, u_*), \sigma] d\sigma$$

are satisfied. Since, by (8.5) and (8.6), the last of these is

$$(8.8) \quad x_*^{n+1}(t_*) = t(t_*) = t(t_{*0}) + \int_{t_{*0}}^{t_*} \mathfrak{N}^*[w, \sigma] d\sigma,$$

and the lower and upper bounds of w on U_* are $\frac{1}{2}$ and $\frac{3}{2}$ respectively, we see that $t(t_*)$ is Lipschitzian, with

$$(8.9) \quad \frac{1}{2} \leq dt/dt_* \leq \frac{3}{2}.$$

Given now a generalized curve $C^* = (\mathfrak{N}^*, x_*(\cdot), [t_{*0}, t_{*1}])$ we define a

corresponding generalized curve

$$(8.10) \quad C = J(C^*)$$

for the original problem. Define

$$(8.11) \quad t_0 = t(t_{*0}), \quad t_1 = t(t_{*1});$$

by (8.9), the function $t(t_*)$ has an inverse function $(t_*(t): t_0 \leq t \leq t_1)$:

$$(8.12) \quad t(t_*(t)) = t, \quad t_0 \leq t \leq t_1.$$

We now define

$$(8.13) \quad y^i(t) = x_*^i(t_*(t)), \quad t_0 \leq t \leq t_1, i = 1, \dots, n,$$

and

$$(8.14) \quad \mathfrak{M}[\phi(u), t] = t_*'(t) \mathfrak{M}^*[w\phi(u), t_*(t)]$$

wherever the right member has a meaning; on the set of measure 0 on which $t_*'(t)$ fails to exist we replace it by, say, $Dt_*'(t)$.

The operator \mathfrak{M} defined by (8.14) clearly has the linearity and positivity requirements of §1; and if $\phi(t, u)$ is continuous and of bounded support on $R \times U$, $w\phi(t, u)$ is continuous and of bounded support on $R \times U^*$, so $\mathfrak{M}[\phi(u, t), t]$ is a measurable function of t . Also, by (8.8) and (8.12),

$$\begin{aligned} \mathfrak{M}[1, t] &= t_*'(t) \mathfrak{M}^*[w, t_*(t)] \\ &= t_*''(t) t'(t_*(t)) \\ &= 1 \end{aligned}$$

for almost all t , so \mathfrak{M} is a relaxed control.

By (8.13), (8.12), (8.7), (8.5), (8.17), (8.14), and (8.13), for $i = 1, \dots, n$,

$$\begin{aligned} (8.15) \quad y^i(t) &= \int_{t_0}^t (dy^i/d\tau) d\tau \\ &= \int_{t_0}^t x_*^{i'}(t_*(t)) t_*'(t) dt \\ &= \int_{t_{*0}}^{t_*^*(t)} x_*^{i'}(\sigma) d\sigma \\ &= \int_{t_{*0}}^{t_*^*(t)} \mathfrak{M}^*[wf^i(x(\sigma), t(\sigma), u), \sigma] d\sigma \\ &= \int_{t_0}^t \mathfrak{M}^*[wf^i(x(t_*(\tau)), \tau, u), t_*(\tau)] (dt_*/d\tau) d\tau \\ &= \int_{t_0}^t \mathfrak{M}[f^i(y(\tau), \tau, u), \tau] d\tau, \end{aligned}$$

so $C = (\mathfrak{M}, y(\cdot), [t_0, t_1])$ is a generalized curve. Since C^* is in \mathfrak{F}^* , $(x_*(t_{*0}), t_{*0}, x_*(t_{*1}), t_{*1})$ is in E_* , so $(x_*(t_{*0}), x_*(t_{*1})) = (y(t_0), t_0, y(t_1), t_1)$ is in E ; therefore C is in \mathfrak{F} , and by (8.4),

$$(8.16) \quad e^*(x_*(t_{*0}), t_{*0}, x_*(t_{*1}), t_{*1}) = e(y(t_0), t_0, y(t_1), t_1) \geq e_{\min}.$$

Now with the \mathfrak{M}_0 and $x_0(\cdot)$ of the minimizing generalized curve C_0 we define a generalized curve C_0^* :

$$(8.17) \quad x_{*0}^i(t_*) = x_0^i(t_*), \quad x_{*0}^{n+1}(t_*) = t_*, \\ t_{0,0} \leq t_* \leq t_{1,0}, \quad i = 1, \dots, n;$$

$$(8.18) \quad \mathfrak{M}_0^*[\phi(u_*), t_*] = \mathfrak{M}_0[\{\phi(u, \frac{1}{2}) + \phi(u, \frac{3}{2})\}/2, t_*].$$

Then C_0^* is easily seen to be a member of \mathfrak{F}^* , and its image, by (8.10), is

$$J(C_0^*) = C_0.$$

In particular, by (8.4),

$$e_*(x_{*0}(t_{*0}), t_{*0}, x_{*0}(t_{*1}), t_{*1}) \\ = e(x_0(t_{0,0}), t_{0,0}, x_0(t_{1,0}), t_{1,0}) \\ = e_{\min},$$

so C^* minimizes e_* on \mathfrak{F}^* . With C_0^* the analogue of (4.2) is verified directly, being merely a notational change of (4.2). Also, hypothesis (5.1) holds for the new problem, so all the conclusions reached in the earlier part of this proof are valid for C_0^* .

We denote by $\lambda_0^*(t_*), \dots, \lambda_{n+1}^*(t_*)$ the multipliers of conclusions (i), (ii), (iii) of Theorem 4.7, and we define

$$\lambda_i(t) = \lambda_i^*(t), \quad i = 0, 1, \dots, n, \quad t_{0,0} \leq t \leq t_{1,0}.$$

Then Theorem 4.7(i) holds, and the analogues of the equations in Theorem 4.7(iii) for $\lambda_1^*, \dots, \lambda_n^*$ imply by (8.18) the equations as written in (iii).

If we define $M^*(x_*, t_*, \lambda^*)$ by the analogue of (4.6), by Theorem 4.7(v) we find that for almost all t in $[t_{0,0}, t_{1,0}]$,

$$(8.19) \quad \sum_{\alpha=1}^{n+1} \lambda_\alpha^*(t) w f^\alpha(x_0(t), t, u) = M^*(x_{*0}(t), t, \lambda^*(t))$$

must hold for all (u, w) in the support of $\mathfrak{M}_0^*[\cdot, t]$. Since the equation $\mathfrak{M}_0[1, t] = 1$ implies that the support of $\mathfrak{M}_0[\cdot, t]$ is not empty, there is a point u_0 in the support. But then, by (8.18), both $(u_0, 1/2)$ and $(u_0, 3/2)$ are in the support of $\mathfrak{M}_0^*[\cdot, t]$, so (8.19) holds for both of these. This is impossible unless the coefficient of w is 0; so

$$(8.20) \quad M^*(x_{*0}(t), t, \lambda^*(t)) = 0 \quad \text{for almost all } t \text{ in } [t_{0,0}, t_{1,0}].$$

By definition, the right member of (8.19) is the infimum of the values of the left member for all u in U , and $f^{n+1} \equiv 1$. Since $w > 0$ on U^* , (8.19) and (8.20) yield for almost all t ,

$$(8.21) \quad \sum_{j=1}^n \lambda_j(t) f^j(x_0(t), t, u) \geq -\lambda_{n+1}^*(t),$$

equality holding if u is in the support of \mathfrak{N}_0 . So for almost all t in $[t_{0,0}, t_{1,0}]$, by (4.6),

$$(8.22) \quad M(x_0(t), t, \lambda(t)) = -\lambda_{n+1}^*(t).$$

Therefore, in the last of the equations in Theorem 4.7(iii) for the new problem (i.e., the equation for λ_{n+1}^*) we may replace λ_{n+1}^* by $-M$ in the left number for almost all t , obtaining the equation in Theorem 4.7 (iv) for such t . If U is compact, the left member of this equation is continuous, so it holds for all t in $[t_{0,0}, t_{1,0}]$. [Otherwise, by (7.11) and the sentence before it, the left member of the equation is upper semicontinuous, so for all t the left member is at least equal to the limit of $M(x_0(t'), t', p(t'))$ as $t' > t$ over the set on which the equation holds, and this limit is the same as the value of the right member at t .] Hence Theorem 4.7(iv) is established.

The functions $\lambda_1, \dots, \lambda_n$ satisfy the homogeneous differential equations Theorem 4.7 (iii), so if $\lambda_0, \lambda_1(t), \dots, \lambda_n(t)$ were all 0 at any point, they would be identically 0. Then at any point t at which equality holds in (8.21), $\lambda_0^*, \dots, \lambda_{n+1}^*(t)$ would all be 0. This is known to be false, so (ii) holds. Since (vi) is merely the transcription of the known inequality for the minimizing generalized curve C_0^* , the proof is complete.

III. FURTHER STUDY OF MINIMIZING GENERALIZED CURVES

9. Another necessary condition. In the preceding chapters we found conditions under which an optimal relaxed-control problem has a solution, and conditions satisfied by such solutions. It is clearly desirable to find conditions under which solutions exist that possess better analytical properties, and the most obvious such property is that of being an ordinary curve, instead of a generalized curve. We shall prove a theorem that under suitable hypotheses guarantees a somewhat better conclusion, namely, that for each generalized curve satisfying the conclusions of Theorem 4.7 (in particular for each minimizing curve), the ordinary curve which is the trajectory of the generalized curve is also admissible and gives the same value to the expression to be minimized.

In preparation for proving this theorem we introduce a name for a set already discussed, and we prove another necessary condition.

For each set of numbers $\lambda_1, \dots, \lambda_n$, each x in B and each real number t , $S(x, t, \lambda)$ is defined to be the set of all u in

- (9.1) U at which $F(x, t, u, \lambda)$ (cf. (4.5)) attains its minimum value $M(x, t, \lambda)$. [When U is not compact, this set may be empty.]

We shall use an abbreviation for a lengthy expression that occurs several times in the following sections.

- (9.2) For all x in B , t real, $(\lambda_0, \dots, \lambda_n)$ in R^{n+1} , u and u_0 in U , we define

$$D(u; x, t, u_0, \lambda) = F_t(x, t, u, \lambda) + \sum_{j=1}^n f^j(x, t, u_0) F_{x^j}(x, t, u, \lambda) \\ - F_t(x, t, u_0, \lambda) - \sum_{j=1}^n F_{x^j}(x, t, u_0, \lambda) f^j(x, t, u).$$

We shall say that

- a mean-value operator \mathfrak{M} satisfies the D_w -condition at (x, t, λ) if the support of \mathfrak{M} is contained in $S(x, t, \lambda)$, and for all u_0 in $S(x, t, \lambda)$,
- (9.3)

$$\mathfrak{M}[D(u; x, t, u_0, \lambda)] = 0$$

is satisfied.

We can now show that every generalized curve satisfying conclusions (i)-(v) of Theorem 4.7 will also satisfy the following condition, which is a weak analogue of the Dresden corner condition [25], [12].

THEOREM 9.4. *Let $C_0 = (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ be a generalized curve satisfying conclusions (i)-(v) of Theorem 4.7 with multipliers $\lambda_1(t), \dots, \lambda_n(t)$. Then for almost all t in $[t_{0,0}, t_{1,0}]$ the mean-value operator $\mathfrak{M}_0[\cdot, t]$ satisfies condition D_w at $(x_0(t), t, \lambda(t))$.*

There is a subset A of $[t_{0,0}, t_{1,0}]$, consisting of almost all t in the interval, such that if t is in A , the equations in Theorem 4.7 (iv), (v) hold and the integrals in (iii), (iv) have derivatives equal to their integrands. Let t_0 belong to A , and let u_0 be in $S(x_0(t_0), t_0, \lambda(t_0))$. The function $F(x_0(t), t, u_0, \lambda(t)) - \mu(t)$ is continuous on the t -interval, and by definition of μ and M (see Theorem 4.7 (vi) and (4.6)) is nonnegative almost everywhere, hence everywhere. At t_0 it vanishes by Theorem 4.7 (v), and since t_0 is in A it is differentiable, so its derivative must be 0. By (1.11), Theorem 4.7 (iii) and (vi), this implies that the equation in (9.3) holds at $(x_0(t_0), t_0, \lambda(t_0))$, and establishes the theorem.

10. Conditions ensuring the existence of an ordinary optimal control.

We can now state and prove the theorem discussed at the beginning of §9.

THEOREM 10.1. *Let the generalized curve $C_0 : (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ satisfy*

conclusions (i)-(v) of Theorem 4.7 with multipliers $\lambda_1(t), \dots, \lambda_n(t)$. Assume that for almost all t in $[t_{0,0}, t_{1,0}]$, to each mean-value operator \mathfrak{M} satisfying condition D_w (cf. (9.3)) at $(x_0(t), t, \lambda(t))$ there is a point $u_1(t)$ in U such that

$$(10.1a) \quad \mathfrak{M}[f^i(x_0(t), t, u), t] = f^i(x_0(t), t, u_1(t)), \quad i = 1, \dots, n.$$

Then there is a measurable function $u_2(t)$, $t_{0,0} \leq t \leq t_{1,0}$, such that

$$(10.1b) \quad \mathfrak{M}_0[f^i(x_0(t), t, u), t] = f^i(x_0(t), t, u_2(t)), \quad i = 1, \dots, n,$$

for almost all t , and $x = x_0(t)$ is a trajectory corresponding to the control $u_2(\cdot)$, and the value of e corresponding to this trajectory is the same as that corresponding to C_0 .

The statement about e is a trivial consequence of the other conclusions, since the value of e corresponding to C_0 is $e(t_{0,0}, x_0(t_{0,0}), t_{1,0}, x_0(t_{1,0}))$, and this is unchanged if we replace C_0 by the ordinary curve $x = x_0(\cdot)$.

In order to prove this we use a generalization [15] of a theorem of Filippov [5]. This generalization has the following corollary.

COROLLARY 10.2. *Let Q be a closed subset of a finite-dimensional space, A_3 a finite-dimensional space, and A_1 a Lebesgue-measurable set. Let $k: Q \rightarrow A_3$ be continuous and $y: A_1 \rightarrow A_3$ measurable. If for each t in A_1 there is a v_1 in Q such that $y(t) = k(v_1)$, then there is a measurable function $v_2(t)$, t in A_1 , with values in Q such that $y(t) = k(v_2(t))$, t in A_1 .*

Let A_3 be R^{2n+1} , and let Q be the set of points (x, t, u) with x in B , t real and u in U ; this is closed. Let A_1 be the set of points t in $[t_{0,0}, t_{1,0}]$ for which (10.1a) is satisfied; then A_1 is almost all of the interval. We define the function $k: Q \rightarrow A_3$ by setting

$$(10.3) \quad \begin{aligned} k^i(x, t, u) &= f^i(x, t, u), & i &= 1, \dots, n, \\ k^i(x, t, u) &= x^{i-n}, & i &= n+1, \dots, 2n, \\ k^{2n+1}(x, t, u) &= t. \end{aligned}$$

Then k is continuous on Q . For all t in A_1 we define

$$(10.4) \quad y^i(t) = \mathfrak{M}_0[k^i(x_0(t), t, u), t].$$

By §1, this is measurable. Now for all t in A_1 , the equations

$$y^i(t) = k^i(x_0(t), t, u_1(t)), \quad i = 1, \dots, 2n+1,$$

are satisfied; the first n are a transcription of (10.1a), and the last $n+1$ are trivial. So, by Corollary 10.2, there is a measurable function $v_2: A_1 \rightarrow Q$ such that

$$(10.5) \quad y^i(t) = k^i(v_2(t)), \quad i = 1, \dots, 2n+1.$$

If we denote the point $v_2(t)$ of Q by the alternative symbol $(x_2(t), t_2(t), u_2(t))$, by (10.3), (10.4) and (10.5) we find for $i = 1, \dots, n$,

$$\mathfrak{N}_0[f^i(x_0(t), t, u), t] = f^i(x_2(t), t_2(t), u_2(t)),$$

$$\mathfrak{N}_0[x_0^i(t), t] = x_2^i(t),$$

$$\mathfrak{N}_0[t, t] = t_2(t).$$

By the last $n + 1$ of these equations, $t_2(t)$ is identically t , and x_2 is identically x_0 ; so the first n equations take the form (10.1b).

It remains only to define $u_2(t)$ to be an arbitrary point of U for t in $[t_{0,0}, t_{1,0}] \setminus A_1$ to obtain a function $u_2(\cdot)$ with all the desired properties.

COROLLARY. *If the $u_1(t)$ of Theorem 10.1 also satisfies*

$$\mathfrak{N}_0[F_{x^i}(x_0(t), t, u, \lambda), t] = F_{x^i}(x_0(t), t, u_1(t), \lambda), \quad i = 1, \dots, n,$$

the λ_i also serve as the multipliers with which the ordinary optimal control $u_2(\cdot)$ satisfies the conclusions of Theorem 4.7.

We need only augment (10.3) by including the F_{x^i} among the components of k .

11. Generalization of the Dresden corner condition. The Dresden "corner condition", for problems of the calculus of variations, is a condition satisfied by the right and left derivatives of a minimizing curve of class D' . Much of the role played by derivatives in classical calculus of variations is taken over by controls in control theory, so we define a relaxed-control analogue for right and left derivatives.

Let $C: (\mathfrak{N}, x(\cdot), [t_0, t_1])$ be a generalized curve, and t' a number such that $t_0 \leq t' < t_1$. Then

$$(11.1) \quad \begin{array}{l} \text{an operator } \mathfrak{N}' \text{ on } C_0[U] \text{ is a right limit of } \mathfrak{N}[\cdot, t] \text{ at } t' \\ \text{provided that there exists a sequence of numbers} \\ t_1 > t_2 > t_3 > \dots \rightarrow t' \text{ such that for every function } \phi \\ \text{in } C_0[U], \end{array}$$

$$(11.1a) \quad \lim_{k \rightarrow \infty} (t_k - t')^{-1} \int_{t'}^{t_k} \mathfrak{N}[\phi(u), \tau] d\tau = \mathfrak{N}'[\phi(u)].$$

Left limits are analogously defined, with $t_0 < t_2 < t_3 < \dots \rightarrow t'$.

If U is bounded, right limits and left limits are mean-value operators, since we may choose $\phi \equiv 1$. [If U is unbounded this may be false; we obtain only $0 \leq \mathfrak{N}'[1] \leq 1$.]

COROLLARY 11.2. *For almost all t' in $[t_0, t_1]$, $\mathfrak{N}[\cdot, t']$ is the unique right limit and the unique left limit of $\mathfrak{N}[\cdot, t]$.*

Let $\phi_1, \phi_2, \phi_3, \dots$ be a countable dense subset of $C_0[U]$. For each k the

function

$$(11.3) \quad \int_{t_0}^t \mathfrak{M}[\phi_k(u), \tau] d\tau, \quad t_0 \leq t \leq t_1,$$

is Lipschitzian, so except on a set N_k of measure 0 its derivative exists and is equal to the integrand. Discarding the union of the N_k leaves a subset A of (t_0, t_1) with measure $t_1 - t_0$ on which (11.3) has derivative equal to its integrand for all k . If \mathfrak{M}' is a right limit of $\mathfrak{M}[\cdot, t]$ at a point t' of A , by (11.1a) we have

$$\mathfrak{M}[\phi_k(u), t'] = \mathfrak{M}'[\phi_k(u)], \quad k = 1, 2, 3, \dots$$

Since the ϕ_k are dense in $C_0[U]$, $\mathfrak{M}[\cdot, t']$ and $\mathfrak{M}'[\cdot]$ are identical.

LEMMA 11.4. *Let $C_0 : (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ satisfy hypotheses (i)-(v) of Theorem 4.7 with multipliers $\lambda_0, \dots, \lambda_n(t)$. If \mathfrak{M}' is a right or left limit of $\mathfrak{M}_0[\cdot, t]$ at t' , the support of \mathfrak{M}' is contained in $S(x_0(t'), t', \lambda(t'))$, and*

$$(11.4a) \quad F(x_0(t'), t', u_0, \lambda(t')) = \mu(t')$$

for all u_0 in the support of \mathfrak{M}' .

[If \mathfrak{M}' has empty support, that is, if $\mathfrak{M}'[\phi] = 0$ for all ϕ , the conclusion is trivial. So we assume that the support is not empty.] To be specific, suppose that \mathfrak{M}' is a right limit. Let A be defined as in the proof of Theorem 9.4, and let t_2, t_3, \dots be as in (11.1). Let u_0 be in the support of \mathfrak{M}' , and let V be any [bounded] neighborhood of u_0 in U . There exists a function $\phi: U \rightarrow R$, nonnegative [and of compact support] on U , vanishing outside V , and positive at u_0 . Then $\mathfrak{M}'[\phi] > 0$, so for all large h , by (11.1),

$$\int_{t'}^{t_h} \mathfrak{M}_0[\phi, \tau] d\tau > 0.$$

Therefore there is a point t'_h of the set A between t and t_h at which $\mathfrak{M}_0[\phi, t'_h] > 0$, and so ϕ must be positive at a point u'_h of the support of $\mathfrak{M}_0[\cdot, t'_h]$. But then u'_h is in V (since $\phi(u'_h) > 0$), and

$$F(x_0(t'_h), t'_h, u'_h, \lambda(t'_h)) - \mu(t'_h) = 0$$

by Theorem 4.7 (v), (vi). We can choose a subsequence (we keep the same notation) of the u'_h that converges to a limit u_v . Then

$$F(x_0(t'), t', u_v, \lambda(t')) - \mu(t') = 0.$$

Since u_v is in the closure of V , we can next choose a sequence of neighborhoods of u_0 contracting to u_0 . Then u_v tends to u_0 , and (11.4a) is satisfied. By Theorem 4.7 (iv), u_0 is in $S(x_0(t'), t', \lambda(t'))$.

LEMMA 11.5. *Let $C: (\mathfrak{M}, x(\cdot), [t_0, t_1])$ be a generalized curve, t' a number such that $t_0 \leq t' < t_1$, and \mathfrak{M}' a right limit of $\mathfrak{M}[\cdot, t]$ at t' (with t_2, t_3, \dots as in*

(11.1)). Let $(\Phi(x, t, u) : x \text{ in } B, t \text{ real}, u \text{ in } U)$ be continuous. [Assume that

(11.5a) there is a set A_0 , consisting of all (x, t) with x in a neighborhood of $x(t')$ in B and t in a neighborhood of t' in $[t_0, t_1]$, and a continuous function $G : U \rightarrow R$ such that 1 and Φ are of slower growth than G uniformly on A_0 , and the following numbers are bounded:

$$(11.5b) \quad (t - t')^{-1} \int_{t'}^t \mathfrak{M}[G(u), \tau] d\tau, \quad t' < t \leq t_1.]$$

Then

$$(11.5c) \quad \mathfrak{M}'[\Phi(x(t'), t', u)] = \lim_{h \rightarrow \infty} (t_h - t')^{-1} \int_{t'}^{t_h} \mathfrak{M}[\Phi(x(\tau), \tau, u), \tau] d\tau.$$

[If U is bounded this is trivial; for then $\Phi(x(\tau), \tau, u)$ tends to $\Phi(x(t'), t', u)$ uniformly on U as $\tau \rightarrow t'$, so the right member of (11.5c) is unaltered if we replace τ by t' in the integrand, and in this form (11.5a) is merely (11.1a).

If U is unbounded, without loss of generality we assume $\Phi \geq 0$. For each positive integer q , with the ρ_q of (1.7) we define

$$\phi_q(x, t, u) = \rho_q(u)\Phi(x, t, u), \quad \psi_q = \Phi - \phi_q;$$

and we let C be an upper bound for the absolute values of the quantities (11.5b). If $\epsilon > 0$, for all large q the inequality

$$|\psi_q(x, t, u)| \leq \epsilon G(t, u)$$

holds for all (x, t) in A_0 and all u in U , by (2.6). Thus the expression

$$(t_h - t')^{-1} \int_{t'}^{t_h} \mathfrak{M}[\phi_q(x(\tau), \tau, u), \tau] d\tau$$

converges to the corresponding expression with Φ in place of ϕ_q , uniformly in h . As $h \rightarrow \infty$, it tends to $\mathfrak{M}'[\phi_q(x(t'), t', u)]$, by the previous proof. Since we can interchange the order of the limit processes,

$$\lim_{q \rightarrow \infty} \mathfrak{M}'[\phi_q(x(t'), t', u)] = \lim_{h \rightarrow \infty} (t_h - t')^{-1} \int_{t'}^{t_h} \mathfrak{M}[\Phi(x(\tau), \tau, u), \tau] d\tau.$$

By definition, the left member of this equation is $\mathfrak{M}'[\Phi(x(t'), t', u)]$; this completes the proof.]

The next theorem is a generalization of the Dresden corner condition.

THEOREM 11.6 Let $C_0 : (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ satisfy conclusions (i)-(v) of Theorem 4.7 with multipliers $\lambda_1(t), \dots, \lambda_n(t)$. Let t' be a number such that $t_{0,0} \leq t' < t_{1,0}$. [Assume that there is a positive number δ and a cube W in R^r such that if $t' < t < \min(t_{1,0}, t' + \delta)$, the support of $\mathfrak{M}_0[\cdot, t]$ is contained

in W .] Then for every right limit \mathfrak{M}' of $\mathfrak{M}_0[\cdot, t]$ at t' and every u_0 in $S(x_0(t'), t', \lambda(t'))$ it is true that

$$(11.6a) \quad \mathfrak{M}'[D(u; x_0(t'), t', u_0, \lambda(t'))] \leq 0.$$

An analogous statement holds for left limits \mathfrak{M}' of $\mathfrak{M}_0[\cdot, t]$ at t' , the left member of (11.5a) being then ≥ 0 .

By Theorem 4.7(iii), if $\lambda_1(t), \dots, \lambda_n(t)$ all vanish at any t , they are identically 0, and the conclusion is trivial. We therefore suppose that they are not all 0 at any t .

Let A be as in the proof of Theorem 9.4. For each t in A we then have, for all u_0 in U ,

$$\mathfrak{M}_0[D(u; x_0(t), t, u_0, \lambda(t)), t] = \frac{d}{dt} [\mu(t) - F(x_0(t), t, u_0, \lambda(t))].$$

The function in square brackets is absolutely continuous, since μ , x_0 and λ are indefinite integrals and F is Lipschitzian on the set of values in question. It is never positive, by Theorem 4.7 (iv), and by (11.4a) it vanishes at t' if u_0 is in $S(x_0(t'), t', \lambda(t'))$. So if $t' < t_h \leq t_{1,0}$,

$$(11.7) \quad \int_{t'}^{t_h} \mathfrak{M}_0[D(u; x_0(\tau), \tau, u_0, \lambda(\tau)), \tau] d\tau \leq 0.$$

For τ in the closed interval $[t', t' + \delta]$ and u in U [if U is bounded, or in $U \cap W$ in the unbounded case], $D(u; x_0(\tau), \tau, u_0, \lambda(\tau))$ is uniformly continuous. Hence

$$(11.8) \quad \begin{aligned} \lim_{t_h \rightarrow t'} (t_h - t')^{-1} \int_{t'}^{t_h} \mathfrak{M}_0[D(u; x_0(t'), t', u_0, \lambda(t')), \tau] d\tau \\ = \lim_{t_h \rightarrow t'} (t_h - t')^{-1} \int_{t'}^{t_h} \mathfrak{M}_0[D(u; x_0(\tau), \tau, u_0, \lambda(\tau)), \tau] d\tau. \end{aligned}$$

By (11.7), the right member of (11.8) is nonpositive. But by proper choice of the t_h , the left member of (11.8) is the left member of (11.6a), establishing the conclusion.

[*Remark.* The hypothesis concerning the cube W can be replaced by the weaker hypothesis that the f^i , etc., are majorized by a function G of faster growth as in the hypotheses of Theorem 4.7, and that the indefinite integral of $\mathfrak{M}_0[G, t]$ has a finite derivative at t' . However, the added generality does not seem to justify the added effort, which all resides in generalizing (11.8).]

12. Right and left derivatives. We now establish a theorem in which both hypotheses and conclusions are stronger than in Theorem 10.1.

THEOREM 12.1. *Let $C_0 : (\mathfrak{M}_0, x_0(\cdot), [t_{0,0}, t_{1,0}])$ satisfy conditions (i)-(v) of Theorem 4.7 with multipliers $\lambda_1(t), \dots, \lambda_n(t)$ not all identically 0. [Assume that there is a cube W in R^r and a positive δ such that if t is in $[t_{0,0}, t_{1,0}]$ and $\|x - x_0(t)\| \leq \delta$ and $\|\lambda - \lambda(t)\| \leq \delta$, the set $S(x, t, \lambda)$ is contained in W .] Assume that to each t' in $[t_{0,0}, t_{1,0}]$ there corresponds a point $u_r(t)$ of*

$S(x_0(t'), t', \lambda(t'))$ with the following properties: a mean value \mathfrak{M}' with support contained in $S(x_0(t'), t', \lambda(t'))$ satisfies

$$(12.1a) \quad \mathfrak{M}'[D(u; x_0(t'), t', u_0, \lambda(t'))] \leq 0$$

for all u_0 in $S(x_0(t'), t', \lambda(t'))$ if and only if

$$(12.1b) \quad \mathfrak{M}'[\phi(u)] = \phi(u_r(t'))$$

for all functions ϕ in $C_0[U]$. Then for each t' such that $t_{0,0} \leq t' < t_{1,0}$, the functions x_0^i , λ_i and μ have right derivatives at t' , and these satisfy

$$(12.1c) \quad \begin{aligned} x_0^{i'}(t'+) &= f^i(x_0(t'), t', u_r(t')), \\ \lambda_i'(t'+) &= -F_{x^i}(x_0(t'), t', u_r(t'), \lambda(t')), \\ \mu'(t'+) &= F_t(x_0(t'), t', u_r(t'), \lambda(t')). \end{aligned}$$

The function $u_r(t)$, $t_{0,0} \leq t \leq t_{1,0}$, is measurable, and is an admissible control. The functions $x_0^i(\cdot)$ are a trajectory corresponding to this control, and the conclusions of Theorem 4.7 are satisfied with this trajectory and the same multipliers λ_i .

Consider first t' such that $t_{0,0} \leq t' < t_{1,0}$. Let t_2, t_3, \dots be a decreasing sequence of points of the interval with limit t' ; let ϕ_1, ϕ_2, \dots be a countable dense subset of $C_0[U]$. By the diagonal process we can select a subsequence t_β , $\beta = \beta_1, \beta_2, \dots$, of t_2, t_3, \dots such that for each k ,

$$\lim_{\beta \rightarrow \infty} (t_\beta - t')^{-1} \int_{t'}^{t_\beta} \mathfrak{M}_0[\phi_k(u), \tau] d\tau$$

exists. The limit then exists with ϕ_k replaced by any ϕ in $C_0[U]$, and it is clear that the limit is a nonnegative linear functional of ϕ . We call it \mathfrak{M}' . Then by (11.1), \mathfrak{M}' is a right limit of $\mathfrak{M}_0[\cdot, t]$. By Lemma 11.4, its support is contained in $S(x_0(t'), t', \lambda(t'))$; by Theorem 11.6, inequality (12.1a) holds. Also, with $\phi \equiv 1$ [on W],

$$\mathfrak{M}'[1] = \lim_{\beta \rightarrow \infty} (t_\beta - t')^{-1} \int_{t'}^{t_\beta} \mathfrak{M}_0[1, \tau] d\tau = 1.$$

Hence by hypothesis, (12.1b) must hold. Now by the same argument as used in proving (11.8),

$$\begin{aligned} f^i(x_0(t'), t', u_r(t')) &= \mathfrak{M}'[f^i(x_0(t'), t', u)] \\ &= \lim_{\beta \rightarrow \infty} (t_\beta - t')^{-1} \int_{t'}^{t_\beta} \mathfrak{M}_0[f^i(x_0(t'), t', u), \tau] d\tau \\ &= \lim_{\beta \rightarrow \infty} (t_\beta - t')^{-1} \int_{t'}^{t_\beta} \mathfrak{M}_0[f^i(x_0(\tau), \tau, u), \tau] d\tau \\ &= \lim_{\beta \rightarrow \infty} (t_\beta - t')^{-1} (x_0^i(t_\beta) - x_0^i(t')). \end{aligned}$$

Since this holds for an arbitrary sequence t_2, t_3, \dots decreasing to limit t' , the right derivative of x_0^i at t' exists and satisfies the first equation in (12.1c). The other parts of (12.1c) are similarly established.

By Theorem 9.4, for almost all t in $[t_{0,0}, t_{1,0}]$ the mean-value operator $\mathfrak{M}_0[\cdot, t]$ satisfies condition D_w , and therefore satisfies (12.1a). Its support is contained in $S(x_0(t), t, \lambda(t))$ by Theorem 4.7 (v), so by hypothesis, for almost all t we have

$$(12.2) \quad \mathfrak{M}_0[\phi(u), t] = \phi(u_r(t))$$

for all ϕ in $C_0[U]$. The limit process by which this is extended to other continuous ϕ is a triviality; (12.2) holds for all continuous ϕ . In particular, if j is one of the numbers $1, \dots, r$ we can take $\phi(u) = \tan^{-1} u^j$. The left member of (12.2) is measurable by (1.10); by (12.2), $\tan^{-1} u_r^j(t)$ is measurable. Hence $u_r(\cdot)$ is a measurable function, and since its values are in U it is an admissible control. The other conclusions follow from this and (12.1c).

COROLLARY 12.3. *If the hypotheses of Theorem 12.1 hold under the changes that the symbol u_r is everywhere replaced by u_l , and the inequality sign in (12.1a) is reversed and t' is assumed to satisfy $t_{0,0} < t' \leq t_{1,0}$, then x^i, μ and λ_i have left derivatives at t' , satisfying the (amended) equations (12.1c).*

The proof is the same apart from obvious changes.

13. Examples and remarks. We now consider an example, essentially the same as that in [3, p. 484], which will serve both to show how Theorem 10.1 applies when enough convexity properties are present, and to show how the apparently more general case in which the domain of the control variable u varies with x and t can be incorporated in the preceding theory.

We shall assume that for each x in R^n and each t in R^1 the set $U(x, t)$ is a closed subset of r -space R^r , and furthermore that the set M of all (x, t, u) with x in R^n , t in R^1 and u in $U(x, t)$ is a closed subset of R^{n+1+r} . As before, E is a closed set in R^{2n+2} , and $f^0(x, t, u), \dots, f^n(x, t, u)$ are defined and continuous for x in R^n , t in R^1 and u in R^r . An (ordinary) curve $u = u(t), x = x(t), t_0 \leq t \leq t_1$, is admissible if $u(t)$ is in $U(x(t), t)$ for almost all t , and (1.1) holds, and the ends $(x(t_0), t_0, x(t_1), t_1)$ are in E . In the class of admissible curves we seek one which minimizes

$$(13.1) \quad e(x(t_0), t_0, x(t_1), t_1) + \int_{t_0}^{t_1} f^0(x(\tau), \tau, u(\tau)) d\tau.$$

To this optimal-control problem there corresponds an optimal relaxed-control problem as in §1. We first assume that it is possible to find a minimizing sequence whose trajectories all lie in a bounded closed subset A of (x, t) -space. We also assume that there is a continuous function $\phi(z)$, $0 \leq z < \infty$, such that $\phi(z)/z \rightarrow \infty$ as $z \rightarrow \infty$ and

$$(13.2) \quad f^0(x, t, u) \geq \phi(|u|)$$

for all (x, t, u) with (x, t) in A and u in $U(x, t)$. Without loss of generality we may assume $\phi \geq 0$. Otherwise, let c be the minimum of ϕ ; if we replace ϕ by $\phi - c$, f^0 by $f^0 - c$ and e by $e + c(t_1 - t_0)$, the quantity (13.1) is unaffected. Also without loss of generality we may assume that (13.2) holds for all x, t , and u . Otherwise, we replace $f^0(x, t, u)$ by $\phi(|u|) + |f^0(x, t, u) - \phi(|u|)|$. This does not decrease (13.1) on any curve, and leaves it unchanged for all admissible curves with trajectories in A . We now add the following assumption.

(13.3) f^1, \dots, f^n are all of slower growth than ϕ uniformly on A .

From the closedness of M it follows that there is a continuous (even an infinitely differentiable) function $h(x, t, u)$, x in R^n , t in R^1 , u in R^r , such that

$$(13.4) \quad 0 \leq h(x, t, u) \leq 1 \quad \text{for all } x, t, u,$$

and

$$(13.5) \quad h(x, t, u) = 0 \quad \text{if and only if } u \text{ is in } U(x, t).$$

We now transform this problem into one of the type previously considered, but in which there are $n + 2$ space variables (which we call y^1, \dots, y^{n+2}) and $r + 1$ control variables (which we call v^1, \dots, v^{r+1}). Since we often wish to return to the lower-dimensional spaces, we use the symbol $x(y)$ to denote the projection (y^1, \dots, y^n) of y , and $u(v)$ to denote (v^1, \dots, v^r) . $V(y, t)$ will denote the set of all points v with $u(v)$ in $U(x(y), t)$. Admissible curves will be those that satisfy the differential equations

$$(13.6) \quad y^i(t) = y^i(t_0) + \int_{t_0}^t g^i(y(\tau), \tau, v(\tau)) d\tau,$$

where we define

$$(13.7) \quad \begin{aligned} g^i(y, t, v) &= f^i(x(y), t, u(y)), & i &= 1, \dots, n, \\ g^{n+1}(y, t, v) &= f^0(x(y), t, u(y)) + [v^{r+1}]^2, \\ g^{n+2}(y, t, v) &= h(x(y), t, u(v)), \end{aligned}$$

and the end conditions

$$(13.8) \quad \begin{aligned} &(x(y(t_0)), t_0, x(y(t_1)), t_1) \text{ in } E, \\ &y^{n+1}(t_0) = y^{n+2}(t_0) = 0, \quad y^{n+2}(t_1) \leq 0. \end{aligned}$$

We seek to minimize

$$(13.9) \quad e(x(y(t_0)), t_0, x(y(t_1)), t_1) + y^{n+1}(t_1)$$

in the class of admissible curves, assuming that class to be nonempty.

If we change this to a relaxed-control problem we find that all the hypotheses of Theorem 2.7 are satisfied; as replacement for $G(u)$ we can use $\phi(|u(v)|) + [v^{r+1}]^2$, and then hypothesis (b) of Theorem 2.7 (iii) holds for $j = n + 1$ and (a) holds for the other values of j . Therefore, by Theorem 2.7, a minimizing generalized curve $(\mathfrak{N}_0, y_0, [t_{0,0}, t_{1,0}])$ exists. Since $g^{n+2} = h \geq 0$, (13.8) implies

$$\mathfrak{N}_0[h(x(y(t)), t, u(v)), t] = 0$$

for almost all t ; for such t , we must have $h(x(y(t)), t, u(v)) = 0$ for all v in the support of \mathfrak{N}_0 , which implies that (v^1, \dots, v^r) is in $U(x(y(t)), t)$ for all such v . Moreover, for almost all t the support of \mathfrak{N}_0 must lie in the hyperplane $v^{r+1} = 0$. Otherwise, we would define \mathfrak{N}' to be the mean-value operator

$$\mathfrak{N}'[\phi(v), t] = \mathfrak{N}_0[\phi(v^1, \dots, v^r, 0), t].$$

This would decrease $y^{n+1}(t_1)$ and the quantity in (13.9) without changing anything else, which is impossible since $(\mathfrak{N}_0, y_0, [t_{0,0}, t_{1,0}])$ minimizes (13.9). We have therefore found a solution of the relaxed-control problem corresponding to the problem stated at the beginning of this section.

This, incidentally, includes as a special case the Bolza problem of minimizing (13.1) (or, rather, its relaxed-control analogue) in a class of arcs satisfying a collection of differential equations and inequalities

$$(13.10) \quad \phi^i(x, t, \dot{x}) = 0, \quad \psi^j(x, t, \dot{x}) \leq 0,$$

if the ϕ^i and ψ^j are continuous. For if $f^i(x, t, u) = u^i, i = 1, \dots, n$, and (13.2) holds, so does (13.3); and if we denote by $U(x, t)$ the set of x with which (13.10) holds, the set M (second paragraph) is closed, so all our hypotheses are satisfied.

It remains to find conditions under which the minimizing generalized curve is an ordinary curve. We define

$$(13.11) \quad \lambda_0 = \lambda_1 = \dots = \lambda_{n+1} = 0, \quad \lambda_{n+2} = 1.$$

Then $F(y, t, v, \lambda) = h(x(y), t, u(v))$, and the set $S(y, t, \lambda)$ on which this attains its minimum is $V(y, t)$. With these multipliers conclusions (i)–(v) of Theorem 4.7 are easily seen to be satisfied. (Note that F_{y^i} and F_t vanish on the support of \mathfrak{N}_0 , since F has its least value 0 at such points.) An almost brutal way of restricting the possibilities for \mathfrak{N}_0 is to add the following hypothesis.

$$(13.12) \quad \text{For all } (x, t), \text{ the set in } R^{n+1} \text{ consisting of all points } (z^0, \dots, z^n) \text{ such that } z^0 \geq f^0(x, t, u), z^i = f^i(x, t, u), \\ i = 1, \dots, n, \text{ for some } u \text{ in } U(x, t) \text{ is a convex set.}$$

This is equivalent to the following.

$$(13.13) \quad \text{For all } (y, t), \text{ the set } Q(y, t) \text{ in } R^{n+1} \text{ consisting of all points } (z^1, \dots, z^{n+1}) \text{ such that } z^i = g^i(y, t, v), \\ i = 1, \dots, n+1, \text{ for some } v \text{ in } V(y, t) \text{ is convex.}$$

Let \mathfrak{M} be a mean-value operator satisfying condition (D_w) (cf. (9.3)) with multipliers (13.11) at a point (y, t) . Then the support of $\mathfrak{M}[\cdot, t]$ is contained in $V(y(t), t)$, and the point with coordinates

$$(\mathfrak{M}[g^1(y(t), t, v), t], \dots, \mathfrak{M}[g^{n+1}(y(t), t, v), t])$$

is the centroid of points in the convex set $Q(y(t), t)$. It therefore is itself in $Q(y(t), t)$, and by definition there is a v_1 in $V(y(t), t)$ such that the equations

$$\mathfrak{M}[g^i(y(t), t, v), t] = g^i(y(t), t, v_1), \quad i = 1, \dots, n+1,$$

are satisfied. This holds also for $i = n+2$, since in this case both members of the equation are 0. This shows that the hypotheses of Theorem 10.1 are satisfied, so there is a minimizing ordinary control yielding the same trajectory $y_0(t)$, or (reverting to the original notation) $x(t)$.

Convexity properties were decisive in the preceding example. We consider another example lacking such properties. We take $n = 2$, $r = 1$, $U = [-1, 1]$, and for simplicity we write u instead of u^1 . We assume that f^1 depends on u only and f^2 on x^1 and u only, both being continuously differentiable for all u in $[-1, 1]$ and all x^1 . The problem is to find $u(t)$ and $x(t)$ satisfying (1.1) and the end conditions

$$(13.14) \quad \begin{aligned} t_0 &= x^1(t_0) = x^2(t_0) = 0, \\ t_1 &\geq 0, \quad x^1(t_1) = 0, \quad x^2(t_1) = 1, \end{aligned}$$

for which t_1 is minimum. We assume that there is at least one curve satisfying these conditions. On any minimizing sequence the values of t_1 are bounded, say $< T$; since $|f^1(u)|$ has a finite upper bound M and x^1 vanishes at 0 and t_1 we have $|x^1(t)| \leq MT/2$, which in turn implies that $|f^2(x^1, u)|$ and $|x^2(t)|$ are bounded. So (2.3) holds, and by Theorem 2.7 there is a minimizing generalized curve $(\mathfrak{M}_0, x_0(\cdot), [0, t_1])$.

It is not necessarily true that an optimal ordinary control exists. For example, Filippov has shown [5] that if

$$(13.15) \quad f^1(x, t, u) = \phi(u), \quad f^2(x, t, u) = \psi(u) + \xi(x^1),$$

where $\phi(u) = u$, $\psi(u) = u^2$ and $\xi(x^1) = -(x^1)^2$, no such control exists. However, we shall now show that the optimal relaxed control is ordinary

provided that the following condition is satisfied:

For all (x, t) and for all distinct points u_0, u_1 of $[-1, 1]$,

$$(13.16a) \quad f_{x^1}^2(x, t, u_0) \neq 0,$$

$$(13.16b) \quad f^1(x, t, u_0)f_{x^1}^2(x, t, u_1) - f^1(x, t, u_1)f_{x^1}^2(x, t, u_0) \neq 0,$$

$$(13.16c) \quad \begin{array}{l} \text{the values of } u \text{ for which } f^1(u) \text{ is maximum are all } < 0 \\ \text{or else are all } > 0, \text{ and likewise for the minima of } f^1(u). \end{array}$$

(This is satisfied in the special case (13.15) if ξ' is nonvanishing and ϕ is strictly monotonic.)

To prove this, let $\lambda_0, \lambda_1(t), \lambda_2(t)$ be multipliers with which conclusions (i)–(vi) of Theorem 4.7 are satisfied. Then

$$(13.17) \quad d\lambda_1/dt = -\lambda_2 \mathfrak{M}_0[f_{x^1}^2(x, t, u), t], \quad d\lambda_2/dt = 0.$$

The constant λ_2 cannot be 0. If it were, λ_1 would also be constant by (13.17). This constant λ_1 could not be 0, for then the minimum value $\mu(t)$ of F would be 0, and by Theorem 4.7 (vi), λ_0 would be 0, contradicting (ii). Nor can λ_1 be a nonzero constant, for then by (13.16c) the set $S(x, t, \lambda)$ at which $F = \lambda_1 f^1(u)$ assumes its minimum is all > 0 , or else is all < 0 , so that x^1 is always > 0 or always < 0 , which is incompatible with (13.14).

The function D of (9.2) simplifies to

$$D(u; x, t, u_0, \lambda) = -\lambda_2 f_{x^1}^2(x, t, u_0) f^1(x, t, u) + \lambda_2 f_{x^1}^2(x, t, u) f^1(x, t, u_0),$$

so by Theorem 9.4 we know that for almost all t , the equation

$$(13.18) \quad -\lambda_2 f_{x^1}^2(x_0(t), t, u_0) \dot{x}^1(t) - \dot{\lambda}_1(t) f^1(x_0(t), t, u_0) = 0$$

holds for all u_0 in $S(x_0(t), t, \lambda)$. Since $\lambda_2 \neq 0$, and (13.16a) holds, $\dot{\lambda}_1(t) \neq 0$. So if u_0, u_1 are both in $S(x_0(t), t, \lambda)$, (13.18) and its analogue with u_1 in place of u_0 have a nontrivial solution. But by (13.16b) this is impossible if $u_0 \neq u_1$. Hence for almost all t the set $S(x_0(t), t, \lambda)$ contains only one point, which we call $u(t)$. The support of \mathfrak{M}_0 is contained in $S(x_0(t), t, \lambda)$, hence is $u(t)$ alone, and for all continuous $\phi(u)$ we have

$$(13.19) \quad \mathfrak{M}_0[\phi(u), t] = \phi(u(t)).$$

So the optimal relaxed control is in fact an ordinary control. (Alternatively, we can refer to Theorem 10.1, since (13.19) implies (10.1a).)

As an example of a conditioned problem in classical calculus of variations we consider the problem of minimizing

$$(13.20) \quad \int_{-1}^1 \{[(\dot{y})^2 - 1]^2 + y\} dt$$

in the class of absolutely continuous curves $y = y(t)$, $-1 \leq t \leq 1$, satis-

fying the end conditions $y(-1) = y(1) = 0$ and the isoperimetric condition

$$(13.21) \quad \int_{-1}^{+1} (\dot{y})^2 dt = b \ (> 0).$$

In control notation, we define

$$\begin{aligned} f^1(x^1, x^2, x^3, t, u) &= u, \\ f^2(x^1, x^2, x^3, t, u) &= u^2, \\ f^3(x^1, x^2, x^3, t, u) &= (u^2 - 1)^2 + x^1, \\ t_0 &= -1, \quad t_1 = 1, \\ x_0 &= (0, 0, 0), \quad x_1^1 = 0, \quad x_1^2 = b, \\ e(t_0, x_0, t_1, x_1) &= x_1^3. \end{aligned}$$

It is easy to see that all curves satisfying (13.21) (i.e., $x_0^2 = 0$, $x_1^2 = b$) lie in a bounded set, say $|x^1| \leq K$, $|x^2| \leq K$, $|x^3| \leq K$. Conditions (4.1) and (4.2) are satisfied if we take $G(u) = u^4 + K$. If we make the trivial change of replacing f^3 by $f^3 + K$ and e by $e - 2K$, the hypotheses of Theorem 2.7 are satisfied, so an optimal generalized curve $C_0 : (\mathfrak{M}_0, x_0(\cdot), [-1, +1])$ exists; it satisfies the conclusions of Theorem 4.7 with multipliers λ_0 , $\lambda_1(t)$, $\lambda_2(t)$, $\lambda_3(t)$. By Theorem 4.7 (iv), $\mu(t)$ is constant, and by (iii),

$$(13.22) \quad \lambda_1' = -\lambda_3, \quad \lambda_2' = 0, \quad \lambda_3' = 0.$$

By (vi),

$$[\lambda_0 - \lambda_3(+1)]\xi_1^3 \geq 0$$

for all ξ_1^3 , so the (constant) λ_3 is equal to λ_0 . If $\lambda_0 = 0$, by (13.22), λ_2 and λ_1 are constants, and $F = \lambda_1 u + \lambda_2 u^2$; for this to attain a minimum anywhere we must have $\lambda_2 > 0$, and then $S(x, t, \lambda)$ consists always of one single point, independent of x and t . Thus x^1 is linear, and since it vanishes at -1 and 1 it is always 0 ; this is incompatible with (13.21). So $\lambda_0 \neq 0$, and we have $\lambda_0 = \lambda_3 = 1$.

Now

$$F = (c - t)u + \lambda_2 u^2 + \{(u^2 - 1)^2 + x^1\}.$$

Hence

$$\begin{aligned} \partial F / \partial u &= (c - t) + 2\lambda_2 u + 4(u^2 - 1)u, \\ \partial^2 F / \partial u^2 &= 2\lambda_2 + 12u^2 - 4. \end{aligned}$$

If $\lambda_2 \geq 2$, this last is nonnegative; $S(x, t, \lambda)$ always consists of one point, and the hypotheses of Theorems 10.1 or 12.1 are trivially satisfied; there is an optimal control which is ordinary and continuous. If $\lambda_2 < 2$, there will be one single value of $t + c$, namely 0, for which F has its minimum at two distinct values, u_1 (> 0) and $-u_1$, of u . For $t > c$, $S(x, t, \lambda)$ contains only one point, which is a value of u greater than u_1 ; for $c - d > t$, $S(x, t, \lambda)$ contains only one point u , and this is $< -u_1$. So the isoperimetric problem has a solution which is an ordinary curve made of two C' arcs meeting at a corner. By choice of c and λ_2 , we fit the end conditions.

REFERENCES

- [1] L. D. BERKOVITZ, *Variational methods in problems of control and programming*, J. Math. Anal. Appl., 3 (1961), pp. 145-169.
- [2] ———, *On control problems with bounded state variables*, Ibid., 5 (1962), pp. 488-498.
- [3] L. CESARI, *Existence theorems for optimal solutions in Lagrange and Pontryagin problems*, this Journal, 3 (1965), pp. 475-498.
- [4] H. G. EGGLESTON, *Convexity*, Cambridge University Press, London, 1958.
- [5] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76-84.
- [6] R. A. GAMBILL, *Generalized curves and the existence of optimal controls*, this Journal, 1 (1963), pp. 246-260.
- [7] E. J. MCSHANE, *Semi-continuity of integrals in the calculus of variations*, Duke Math. J., 2 (1936), pp. 597-616.
- [8] ———, *A navigation problem in the calculus of variations*, Amer. J. Math., 59 (1937), pp. 327-334.
- [9] ———, *Some existence theorems in the calculus of variations, III. Existence theorems for non-regular problems*, Trans. Amer. Math. Soc., 45 (1939), pp. 151-171.
- [10] ———, *On multipliers for Lagrange problems*, Amer. Math., 61 (1939), pp. 809-819.
- [11] ———, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513-536.
- [12] ———, *Necessary conditions for generalized-curve problems of the calculus of variations*, Ibid., 7 (1940), pp. 1-27.
- [13] ———, *Existence theorems for Bolza problems in the calculus of variations*, Ibid., 7 (1940), pp. 28-61.
- [14] ———, *Integration*, Princeton University Press, Princeton, 1944.
- [15] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41-47.
- [16] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [17] L. TONELLI, *Fondamenti di Calcolo delle Variazione*, Zanichelli, Bologna, 1923.
- [18] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111-128.
- [19] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129-145.
- [20] ———, *Minimizing variational curves restricted to a preassigned set*, Trans. Amer. Math. Soc., 112 (1964), pp. 432-455.

- [21] ———, *Variational problems with unbounded controls*, this Journal, 3 (1965), pp. 424–438.
- [22] L. C. YOUNG, *On approximation by polygons in the calculus of variations*, Proc. Royal Soc. Ser. A, 141 (1933), pp. 325–341.
- [23] ———, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. et Lettres, Varsovie, Cl. III, 30 (1937), pp. 212–234.
- [24] ———, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239–258.
- [25] A. DRESDEN, *The second derivatives of the extremal integral*, Trans. Amer. Math. Soc., 9 (1908), pp. 467–486.

OPTIMAL STATIONARY CONTROL OF A LINEAR SYSTEM WITH STATE-DEPENDENT NOISE*

W. M. WONHAM†

1. Introduction. Consider the linear control system described by the formal, vector stochastic differential equation

$$(1.1) \quad \dot{x} = Ax - Bu + C\dot{w}_1 + G(x)\dot{w}_2.$$

In (1.1), u is the control and \dot{w}_1 , \dot{w}_2 are independent Gaussian white noise disturbances.¹ The elements of the matrix G are assumed to be linear in x ; and so the term $G(x)\dot{w}_2$ represents a disturbance of which the intensity is roughly proportional to the deviation of x from the origin $x = 0$. Equivalently, the disturbance can be regarded as a wideband random perturbation of the system matrix A .

Now consider the problem of choosing a feedback control $u = \phi(x)$ such that, in the steady state, the expected quadratic cost

$$(1.2) \quad \varepsilon\{x'Mx + u'Nu\}$$

is a minimum. If $G(x) \equiv 0$, the solution of this problem is well known [1], [2]. Under mild restrictions the optimal control always exists and is a linear function of x which is independent of the intensity of the additive disturbance $C\dot{w}_1$. In the present article it is shown that an optimal control exists for the more general system (1.1), provided the state-dependent noise $G(x)\dot{w}_2$ is sufficiently small. The optimal control is again linear, but is now rather critically dependent on the coefficients of G . Examples are provided to show that instability may result if this dependence is ignored.

The problem is stated precisely in §2; the proof of existence is given in §§3 and 4; and some examples studied in §§5 and 6. We conclude with some remarks on the interpretation of (1.1) and discuss alternative optimization problems which are closely related.

2. Statement of the problem. To make (1.1) precise we assume that x is an n -vector with stochastic differential

$$(2.1) \quad dx = Ax dt - Bu dt + C dw_1 + G(x) dw_2.$$

* Received by the editors June 15, 1966, and in revised form November 7, 1966.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. This research was supported in part by the National Aeronautics and Space Administration under Grant NGR 40-002-015, in part by the Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR-693-66, and in part by the National Science Foundation under Grant GK-967.

This paper will also appear in Proceedings of the First All-Union Symposium on Statistical Problems in Technical Cybernetics (in Russian), Moscow, 1967.

¹ A precise interpretation of (1.1) is given in §2.

Here and below, all vectors and matrices have real-valued elements; A , B and C are constant matrices of dimensions $n \times n$, $n \times m$ and $n \times d_1$ respectively; $G(x)$ is an $n \times d_2$ matrix given by

$$(2.2) \quad G(x) = \sum_{k=1}^n x_k G_k,$$

where the $n \times d_2$ matrices G_k are constants. It is assumed that (A, B) is controllable, and that CC' is positive definite: that is, $d_1 \geq n$ and C is of rank n . The latter assumption obviates fussy discussion about possible degeneracy of the ergodic measure introduced below. Finally, w_1 and w_2 are independent Wiener processes of dimension d_1 , d_2 respectively.

In the following, E denotes Euclidean n -space; a prime ($'$), the transpose of a vector or matrix; and $|\cdot|$, the Euclidean norm: for a matrix F , $|F| = \max \{|Fx| : |x| = 1\}$.

In (2.1) let $u = \phi(x)$, where ϕ is defined on E and satisfies a uniform Lipschitz condition

$$(2.3) \quad |\phi(x) - \phi(y)| \leq k |x - y|, \quad x, y \in E.$$

With this choice of u , (2.1) becomes a stochastic differential equation of Itô's type [3]:

$$(2.4) \quad dx(t) = Ax(t)dt - B\phi[x(t)]dt + Cdw_1(t) + G[x(t)]dw_2(t).$$

If $x(0)$ is a random variable independent of the w_1 , w_2 increments, then (2.4), defined for $t \geq 0$, determines a diffusion process

$$X_\phi = \{x_\phi(t) : t \geq 0\}.$$

Diffusion processes are discussed extensively in [4]; a brief summary can be found in [5].

Of interest here is the case when X_ϕ is positive recurrent (for the definition of this term, see [5]). Under this condition it is known that there exists a unique ergodic probability measure μ_ϕ defined on the Borel sets of E : that is, if the distribution of $x(0)$ is μ_ϕ , then so is that of $x(t)$ for all $t > 0$. Let Φ be the class of *admissible* control functions ϕ , with the properties:

- (i) ϕ satisfies (2.3) for some constant k ;
- (ii) X_ϕ is ergodic;
- (iii) the corresponding ergodic measure μ_ϕ is such that

$$(2.5) \quad \varepsilon_\phi\{|x|^2\} \equiv \int_E |x|^2 \mu_\phi(dx) < \infty.$$

Now define

$$(2.6) \quad L(x, u) = x'Mx + u'Nu,$$

where M , N are constant symmetric positive definite matrices of dimensions $n \times n$, $m \times m$ respectively.

Our problem is the following: find a control $\phi^0 \in \Phi$ which is optimal in the sense that

$$\mathcal{E}_{\phi^0}\{L(x, \phi^0)\} = \min [\mathcal{E}_{\phi}\{L(x, \phi)\} : \phi \in \Phi].$$

3. Existence of an admissible control. In this section it will be shown that Φ is nonempty provided the matrices G_k of (2.2) are sufficiently small. This result will follow from the stability theorem stated below.

Let $V = V(x)$ be of class $C^{(2)}$ on E and let \mathcal{L}_u denote the elliptic operator

$$(3.1) \quad \mathcal{L}_u V(x) \equiv \frac{1}{2} \operatorname{tr} \{C' V_{xx}(x) C + G(x)' V_{xx}(x) G(x)\} \\ + (Ax - Bu)' V_x(x).$$

In (3.1), tr denotes trace, V_x the vector $[\partial V / \partial x_i]$ and V_{xx} the matrix $[\partial^2 V / \partial x_i \partial x_j]$. The operator \mathcal{L}_{ϕ} , obtained by setting $u = \phi(x)$ in (3.1), is the *differential generator* of X_{ϕ} (see [4]).

The following theorem is an immediate consequence of (2.6) and Theorem 4.1 of [6].

THEOREM 3.1. *Let $\phi(x)$ satisfy (2.3). If there exist a function $V(x)$ of class $C^{(2)}$ on E , and a positive number λ such that*

$$(3.2) \quad V(x) \rightarrow \infty \quad \text{as } |x| \rightarrow \infty,$$

and

$$(3.3) \quad \mathcal{L}_{\phi} V(x) \leq \lambda - L[x, \phi(x)], \quad x \in E,$$

then $\phi \in \Phi$.

To apply the theorem, set

$$(3.4a) \quad \phi(x) = Kx,$$

$$(3.4b) \quad V(x) = x' P x,$$

where K, P are constant $m \times n, n \times n$ (respectively) matrices, to be determined so that

$$(3.5) \quad \mathcal{L}_{\phi} V(x) = \lambda - L[x, \phi(x)], \quad x \in E.$$

Let $\Gamma(P)$ be the symmetric $n \times n$ matrix with elements

$$(3.6) \quad [\Gamma(P)]_{kl} = \operatorname{tr} (G_k' P G_l).$$

Then a brief calculation shows that (3.4) determines a solution of (3.5) if and only if

$$(3.7) \quad \lambda = \operatorname{tr} (C' P C)$$

and

$$(3.8) \quad \Gamma(P) + (A - BK)' P + P(A - BK) + M + K' N K = 0.$$

By our assumption of controllability, K can be chosen so that all eigenvalues of the matrix $A - BK$ have negative real parts.² With K so chosen, the following lemma shows that (3.8) has a unique positive definite solution P provided $\sum_k |G_k|^2$ is sufficiently small. This together with Theorem 3.1 implies that $\phi \in \Phi$.

LEMMA 3.1. *If $Q > 0$ and A is stable, the equation*

$$(3.9) \quad \Gamma(P) + A'P + PA + Q = 0$$

has a unique solution $P > 0$ provided

$$(3.10) \quad d_2 \left(\sum_{k=1}^n |G_k|^2 \right) \left| \int_0^\infty e^{tA'} e^{tA} dt \right| < 1.$$

Here and below $P > 0$ (≥ 0) means P is positive definite (semi-definite); $P_1 > P_2$ means $P_1 - P_2 > 0$, etc.

Proof. Equation (3.9) is equivalent to the equation

$$(3.11) \quad P = R + T(P),$$

where

$$R = \int_0^\infty e^{tA'} Q e^{tA} dt$$

and

$$(3.12) \quad T(P) = \int_0^\infty e^{tA'} \Gamma(P) e^{tA} dt.$$

We observe that $\Gamma(P)$ is a linear function of P and $\Gamma(P) \geq 0$ if $P \geq 0$; it follows that $T(P)$ has the same properties. Define

$$P_1 = R, \quad P_{\nu+1} = R + T(P_\nu), \quad \nu = 1, 2, \dots$$

The sequence P_ν is monotone nondecreasing; it is bounded if, for some $\theta \in (0, 1)$ and all $P \geq 0$,

$$(3.13) \quad T(P) \leq \theta |P| I.$$

If (3.13) holds, it follows by a result on positive operators (e.g., [7, Theorem 1, p. 189]) that the matrix

$$P = \lim_{\nu \rightarrow \infty} P_\nu$$

exists; and $P \geq R > 0$. If (3.10) holds, T is actually a contraction; thus (3.13) holds, and P is unique. The proof is complete.

² This fact is easily proved by inspection of the canonical form for (A, B) obtained in [14].

Define

$$\kappa = \inf_K \left| \int_0^\infty e^{t(A-BK)'} e^{t(A-BK)} dt \right|.$$

Thus, for some K , (3.8) has a (unique) solution $P > 0$ provided

$$(3.14) \quad \sum_{k=1}^n |G_k|^2 < (\kappa d_2)^{-1}.$$

4. Existence of an optimal control. It will be shown that an optimal control ϕ^0 exists whenever (3.14) holds, and that ϕ^0 is linear. We use dynamic programming and the well-known method of approximation in policy space [8]. This approach was suggested by the work of Howard, who studied a similar problem for Markov chains [9]. The result depends on the following optimality theorem.

THEOREM 4.1. *Suppose there exist $\phi^0 \in \Phi$, a function $v(x)$ of class $C^{(2)}$ on E , and a positive number λ , with the following properties:*

$$(4.1) \quad \mathcal{E}_\phi \{ |v(x)| + |x| |v_x(x)| + |x|^2 |v_{xx}(x)| \} < \infty \quad \text{for every } \phi \in \Phi;$$

$$(4.2) \quad \mathcal{L}_{\phi^0} v(x) + L[x, \phi^0(x)] = \lambda, \quad x \in E;$$

$$(4.3) \quad \mathcal{L}_u v(x) + L(x, u) \geq \lambda \quad \text{for every } m\text{-vector } u, \quad x \in E.$$

Then ϕ^0 is optimal. Furthermore,

$$(4.4) \quad \lambda = \mathcal{E}_{\phi^0} \{ L(x, \phi^0) \}.$$

Combining (4.2) and (4.3) we obtain the appropriate version of Bellman's equation:

$$(4.5) \quad \min_u \{ \mathcal{L}_u v(x) + L(x, u) \} = \lambda.$$

To prove Theorem 4.1 we need the following lemma.

LEMMA 4.1. *Let X be a diffusion process determined by (2.4), with differential generator \mathcal{L} and ergodic measure μ . If $v(x)$ is a function of class $C^{(2)}$ such that*

$$\mathcal{E}_\mu \{ |v(x)| + |x| |v_x(x)| + |x|^2 |v_{xx}(x)| \} < \infty,$$

then

$$\mathcal{E}_\mu \{ \mathcal{L}v(x) \} = 0.$$

A proof is given in Appendix 1.

To prove Theorem 4.1 observe that if $\phi \in \Phi$ then, by (4.2) and (4.3),

$$\lambda \leq \mathcal{L}_\phi v(x) + L[x, \phi(x)], \quad x \in E.$$

Taking expectations with respect to μ_ϕ on both sides, and applying Lemma 4.1, we obtain

$$\lambda \leq \mathcal{E}_\phi\{L(x, \phi)\}.$$

Again by Lemma 4.1, (4.2) implies

$$\lambda = \mathcal{E}_{\phi^0}\{L(x, \phi^0)\},$$

and the result follows.

To compute an optimal control we seek a solution of Bellman's equation, in the form

$$(4.6) \quad v(x) = x'Px.$$

Substitution shows that (4.5) holds if and only if P satisfies (3.8), with

$$(4.7) \quad K = N^{-1}B'P.$$

The control determined by (4.5) is

$$(4.8) \quad \phi^0(x) = Kx.$$

We show next that (3.8) and (4.7) can be solved for a unique positive definite matrix P . For $\nu = 1, 2, \dots$, let P_ν be a solution of (3.8) with $K = K_\nu$, and define

$$(4.9) \quad K_{\nu+1} = N^{-1}B'P_\nu.$$

By Lemma 3.1, we can choose K_1 so that P_1 exists. It will be shown that if K_2 is defined by (4.9), then P_2 exists and $0 < P_2 \leq P_1$. Write $v_\nu(x) = x'P_\nu x$, $\phi_\nu = K_\nu x$ and $\mathcal{L}_\nu = \mathcal{L}_{\phi(\nu)}$. It can be verified directly that (4.9) is equivalent to the condition

$$(4.10) \quad \mathcal{L}_{\nu+1}v_\nu(x) + L[x, \phi_{\nu+1}(x)] \leq \mathcal{L}_\nu v_\nu(x) + L(x, u), \quad x \in E,$$

for all m -vectors u . That is, $\phi_{\nu+1}$ is determined by the minimizing operation (4.5) applied to v_ν . Setting $\nu = 1$ and $u = \phi_1(x)$ in (4.10), and using (3.8), we see that

$$(4.11) \quad \begin{aligned} -Q &\equiv \Gamma(P_1) + (A - BK_2)'P_1 + P_1(A - BK_2) + M + K_2'NK_2 \\ &\leq 0. \end{aligned}$$

Write $A_2 = A - BK_2$. Since $P_1 > 0$ satisfies (4.11), it follows (by a standard Liapunov theorem) that A_2 is stable. Hence,

$$(4.12) \quad P_1 = \int_0^\infty e^{tA_2'}[M + K_2'NK_2 + \Gamma(P_1) + Q]e^{tA_2} dt.$$

Now P_2 is to be determined by (3.8) with $K = K_2$, or

$$(4.13) \quad P_2 = \int_0^\infty e^{tA_2'} [M + K_2' N K_2 + \Gamma(P_2)] e^{tA_2} dt.$$

As in the proof of Lemma 3.1, we solve (4.13) by successive approximations. Setting $P_2^{(1)} = 0$, we have

$$\begin{aligned} P_2^{(2)} &= \int_0^\infty e^{tA_2'} (M + K_2' N K_2) e^{tA_2} dt \\ &\leq P_1; \end{aligned}$$

and similarly $P_2^{(\kappa)} \leq P_1$, $\kappa = 2, 3, \dots$. Since the $P_2^{(\kappa)}$ are nondecreasing and bounded,

$$(4.14) \quad P_2 \equiv \lim_{\kappa \rightarrow \infty} P_2^{(\kappa)}$$

exists and satisfies (4.13). Thus $P_2 \leq P_1$, and $M > 0$ implies $P_2 > 0$.

It is not asserted that the solution of (4.13) is unique; however, we may now proceed by induction and define

$$P_\nu = \lim_{\kappa \rightarrow \infty} P_\nu^{(\kappa)}, \quad \nu = 1, 2, \dots$$

In this way we obtain a sequence $\{P_\nu\}$ with $0 < P_{\nu+1} \leq P_\nu$. Then

$$(4.15) \quad \begin{aligned} P &= \lim_{\nu \rightarrow \infty} P_\nu, \\ K &= N^{-1} B' P \end{aligned}$$

exist and satisfy (3.8) and (4.7).

Define

$$(4.16) \quad \begin{aligned} \phi^0(x) &= Kx, \\ v(x) &= x'Px, \\ \lambda &= \text{tr}(C'PC). \end{aligned}$$

Theorem 4.1 will be applied to show that ϕ^0 is optimal. By construction, ϕ^0 satisfies (4.2) and (4.3). Furthermore, if $\phi \in \Phi$ then (2.5) and (4.16) imply the truth of (4.1). The existence of ϕ^0 is now established.

We observe that ϕ^0 is unique in the class of linear controls; for if ϕ is another optimal linear control and $\hat{\lambda}$, \hat{P} are the corresponding quantities determined as before, then by (4.4), $\hat{\lambda} = \lambda$, and by (4.16),

$$(4.17) \quad \text{tr}(C'\hat{P}C) = \text{tr}(C'PC).$$

Since \hat{P} , P are independent of C , (4.17) holds for all C , and from this it easily follows that $\hat{P} = P$. Uniqueness of ϕ is a consequence of (4.7).

5. Example 1. The following artificial example is of interest because it illustrates the qualitative dependence of the control law on the intensity of the state-dependent noise. Let

$$(5.1) \quad dx_i = ax_i dt - bu_i dt + c dw_{1i} + g |x| dw_{2i}, \quad i = 1, \dots, n,$$

and

$$L(x, u) = |x|^2 + |u|^2.$$

In (5.1) the matrix $G(x) = g |x| I$ is not linear in x (cf. (2.2)); nevertheless, because of the rotational symmetry, the methods used above apply equally well here, and (3.8), (4.7) become

$$\begin{aligned} g^2(\text{tr } P)I + (aI - bK)'P + P(aI - bK) + I + K'K &= 0, \\ K &= bP. \end{aligned}$$

This gives $P = pI$, $K = bpI$, and $\lambda = nc^2p$, where

$$(5.2) \quad \begin{aligned} p &= (2b^2)^{-1}\{2a + ng^2 + [(2a + ng^2)^2 + 4b^2]^{1/2}\} \\ &\sim nb^{-2}g^2, \end{aligned} \quad g \rightarrow \infty.$$

For large g , $\phi^0(x) \sim nb^{-1}g^2x$, and the optimal control depends rather critically on noise intensity.

Now suppose that for some k , $u = \phi(x) = kx$ in (5.1). Solution of (3.5) and application of (4.7) yield

$$\begin{aligned} \lambda_\phi &= \varepsilon_\phi\{|x|^2 + |u|^2\} \\ &= nc^2(1 + k^2)[2(bk - a) - ng^2]^{-1} \end{aligned}$$

provided

$$(5.3) \quad bk - a > ng^2/2.$$

If this inequality fails (i.e., control is not sufficiently vigorous), then instability results, in the sense that either $\lambda_\phi = +\infty$ or λ_ϕ is not defined: that is, X_ϕ is no longer ergodic. Using the methods of [5] one can show that X_ϕ is ergodic (i.e., μ_ϕ exists) if and only if $bk - a > (n - 2)g^2/2$.

6. Example 2. State-dependent noise and instability. The example of this section illustrates the fact that an admissible linear control need not exist if the intensity of state-dependent noise is sufficiently large. Let

$$(6.1) \quad A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad G(x) = \sqrt{\gamma} \begin{pmatrix} x_1 & x_2 \\ -x_2 & x_1 \end{pmatrix},$$

where $\gamma > 0$ is a constant; thus $\Gamma(P) = \gamma(\text{tr } P)I$.

We first show that, for the present example, (3.8) and (4.7) have a positive solution P if and only if $\gamma < 1$. Let $K = (k_1, k_2)$; then $A - BK$ is

stable if and only if $k_1 > 0$, $k_2 > 0$. If (3.8) has a solution $P > 0$, then $A - BK$ is stable; thus

$$(6.2) \quad P = R + T(P),$$

where

$$R = \int_0^\infty e^{t(A-BK)'} (M + K'NK) e^{t(A-BK)} dt$$

and

$$(6.3) \quad T(P) = \gamma(\operatorname{tr} P) \int_0^\infty e^{t(A-BK)'} e^{t(A-BK)} dt.$$

Denote the integral in (6.3) by S . Computation yields

$$\inf(\operatorname{tr} S; k_1 > 0, k_2 > 0) = 1,$$

and the infimum is not attainable. Then

$$\begin{aligned} P &\geq T(P) \geq T^{(\nu)}(P) \\ &= \gamma^\nu(\operatorname{tr} P)(\operatorname{tr} S)^{\nu-1} S \\ &> \gamma^\nu(\operatorname{tr} P) S, \end{aligned} \quad \nu = 2, 3, \dots,$$

and necessarily $\gamma < 1$.

On the other hand, if $\gamma < 1$ there exist $k_1 > 0$, $k_2 > 0$ such that $\operatorname{tr} S < \gamma^{-1}$, and then

$$P = \sum_{\nu=0}^{\infty} T^{(\nu)}(R) > 0$$

exists and satisfies (3.8). Thus the construction used in §4 succeeds if and only if $\gamma < 1$.

Next we show that no admissible linear control exists if $\gamma > 1$. For this we need the following instability theorem, the proof of which is given in Appendix 2. Let X be a diffusion process determined by (2.4), with differential generator \mathfrak{L} ; let $L(x) \geq 0$, and Hölder continuous, $x \in E$; and let $V_1(x)$, $V_2(x)$ be a pair of real-valued functions with the properties:

- (i) for some $r < \infty$, V_1 , V_2 are defined and of class $C^{(2)}$ for $|x| > r$;
- (ii) there is a sequence $\{x_n\}$ with $|x_n| \rightarrow \infty$ such that $V_1(x_n) \rightarrow +\infty$;
- (iii) $V_2(x) > 0$, $|x| > r$;
- (iv) $\overline{\lim}_{\rho \rightarrow \infty} \frac{\max \{V_1(x) : |x| = \rho\}}{\min \{V_2(x) : |x| = \rho\}} = 0$;
- (v) $\mathfrak{L}V_1(x) \geq 0$, $\mathfrak{L}V_2(x) \leq L(x)$, $|x| > r$.

THEOREM 6.1 (Instability). *Suppose X is positive recurrent, with ergodic probability measure μ . If there exist functions V_1 , V_2 with properties (i)–(v),*

then

$$\varepsilon_\mu\{L(x)\} = +\infty.$$

To apply the theorem in the present case, let $L(x) = |x|^2$ and suppose $\gamma > 1$. We choose $V_2(x) = \theta |x|^2$ for suitable θ , $0 < \theta < 1$; clearly this is possible. More critical is the choice of V_1 : let

$$V_1(x) = [Q(x)]^q,$$

where $Q(x) = Q(x_1, x_2)$ is a positive definite quadratic form, and $0 < q < 1$. To find suitable Q and q , observe that

$$\mathcal{L}V(x) \equiv \frac{1}{2}\gamma |x|^2 \operatorname{tr} V_{xx} + \frac{1}{2} \operatorname{tr} (C'V_{xx}C) + x'(A - BK)'V_x.$$

Suppose first that $A - BK$ is stable. We choose Q so that

$$x'(A - BK)'Q_x(x) = -|x|^2.$$

Write $Q(1, 0) + Q(0, 1) = \operatorname{tr} Q$. Computation yields $\inf \{ \operatorname{tr} Q : k_1 > 0, k_2 > 0 \} = 1$, and

$$\begin{aligned} \mathcal{L}V_1(x) = q |x|^2 Q(x)^{q-1} & \left[\frac{1}{2} \gamma (q-1) \frac{|Q_x(x)|^2}{Q(x)} + \gamma \operatorname{tr} Q - 1 \right] \\ & + O(|x|^{2q-2}), \quad |x| \rightarrow \infty. \end{aligned}$$

Q being a positive quadratic form, we find

$$\sup_x \frac{|Q_x(x)|^2}{Q(x)} \leq 4 \operatorname{tr} Q,$$

so that $\mathcal{L}V_1(x) \geq 0$ for $|x|$ sufficiently large if

$$\gamma(\operatorname{tr} Q)\{2(q-1) + 1\} - 1 > 0.$$

Setting $q = (3 + \gamma^{-1})/4$, we have satisfied all the conditions of Theorem 6.1.

Finally, suppose $A - BK$ is not stable. Then integration of the stochastic differential equation

$$dx = (A - BK)x dt + C dw_1 + G(x) dw_2$$

yields

$$\varepsilon\{|x(t)|^2\} \geq \operatorname{tr} \left\{ \int_0^t e^{s(A-BK)} C C' e^{s(A-BK)'} ds \right\} \rightarrow \infty, \quad t \rightarrow \infty;$$

and the control $\phi(x) = Kx$ cannot be admissible for any value of γ .

To conclude this section we remark that the boundary case $\gamma = 1$ apparently presents difficulty to application of Theorem 6.1, and will not be

discussed at present. Our purpose has been to show that linearly state-dependent noise of sufficient intensity may make stabilization by linear control impossible, regardless of the choice of control parameters, even if the pair (A, B) is controllable in the usual sense. Although this result is not surprising, the situation deserves more detailed study than will be given here.

7. An alternative interpretation of (1.1). It is worth emphasizing that the choice of Itô's equation (2.1) as a precise version of (1.1) is somewhat arbitrary. We shall discuss briefly an alternative version of (1.1) which may be more appropriate in engineering applications. Equation (1.1) is a purely formal equation since the "derivatives" \dot{w}_1, \dot{w}_2 do not exist. In writing (1.1), we usually have in mind a physical system perturbed by noise with a power spectral density which is essentially constant within the frequency passband of the system. However, total noise power is presumably finite, and this fact is overlooked in adopting the precise model (2.1). Thus the question arises whether the diffusion process determined by (2.1) adequately reflects the properties of the physical random process of which (1.1) is a rough description. This question has been discussed in a precise fashion by Stratonovich [10], [11], and by Wong and Zakai [12]. It turns out that the proper Itô equation to associate with (1.1) will depend on what definition is adopted of the formal stochastic integral

$$(7.1) \quad J = \int_a^b G[x(t)] \dot{w}(t) dt.$$

Let $\{t_\nu\}$ be a partition of the interval $[a, b]$. On the basis of results of [10]–[12] it is natural to adopt for (7.1) the definition

$$J = \text{l.i.m.} \sum_\nu G \left[\frac{x(t_\nu) + x(t_{\nu+1})}{2} \right] [w(t_{\nu+1}) - w(t_\nu)]$$

as $\max_\nu (t_{\nu+1} - t_\nu) \rightarrow 0$. Let us now suppose that $x(t)$ has the Itô stochastic differential

$$dx(t) = f(x) dt + G(x) dw,$$

where $G(x) = [g_{ij}(x)]$. Then it can be shown [10] that

$$(7.2) \quad J = \frac{1}{2} \int_a^b G_x[x(t)] \cdot G[x(t)] dt + \int_a^b G[x(t)] dw(t),$$

where the second integral in (7.2) is an Itô stochastic integral, and $G_x \cdot G$ is the vector with i th component

$$\sum_{j,k} (\partial g_{ij} / \partial x_k) g_{kj}.$$

This result means that an alternative natural interpretation of (1.1) is that the process $x(t)$ has the Itô stochastic differential

$$(7.3) \quad dx = [Ax - Bu + \frac{1}{2}G_x(x) \cdot G(x)] dt + C dw_1 + G(x) dw_2.$$

Equation (7.3) differs from (2.1) by the presence of an additional drift term contributed by the coefficient of the state-dependent noise.

Suppose that $G(x)$ has the form (2.2). Then (7.3) can be written

$$dx = \hat{A}x dt - Bu dt + C dw_1 + G(x) dw_2,$$

where \hat{A} is a modified system matrix with elements

$$\hat{a}_{ij} = a_{ij} + \frac{1}{2} \sum_{k,l} g_{ikl} g_{lkj}$$

and g_{ikl} is the (i, k) th element of G_l . After this modification the discussion of §§2-5 remains unchanged.

In light of this discussion consider again Example 1. Here $G(x) = g|x|I$, and

$$G_x(x) \cdot G(x) = g^2 x.$$

Thus $\hat{A} = aI + (g^2/2)I$, and the previous results hold with this replacement. With the new model,

$$\phi^0(x) \sim (n+1)b^{-1}g^2x, \quad g \rightarrow \infty;$$

that is, the optimal control gain is somewhat higher than previously. Suppose next that $u = \phi(x) = kx$. Then (cf. §5) $\lambda_\phi < \infty$ if and only if

$$bk - a > (n+1)g^2/2.$$

Comparing this result with (5.3) we see that the choice of mathematical model may be critical in an assessment of the stability properties of the physical system of interest.

8. Alternative problems. A variety of linear regulator problems with linearly state-dependent noise can be discussed by methods similar to the foregoing. If the index of performance is expectation of a quadratic functional, and if no a priori bound is placed on magnitude of the control vector, then in general the optimal control (when it exists) is linear in x and depends on noise intensity.

To mention one interesting variant, let

$$(8.1) \quad dx = Ax dt - Bu dt + G(x) dw,$$

and consider the problem of minimizing

$$(8.2) \quad \mathcal{E}_x \left\{ \int_0^\infty [x(t)' M x(t) + u(t)' N u(t)] dt \right\}.$$

If $u = \phi(x)$ and $\phi(0) = 0$, then (8.1) admits the null solution $x(t) \equiv 0$ (see, e.g., [13]). The functional (8.2) is finite provided $x = 0$ is globally asymptotically stable in an appropriate sense. By a slight extension of the methods of [13], one can show that X_ϕ is stable if and only if a continuous

function $V(x)$ exists such that:

- (i) $V(x) > 0$, $x \neq 0$, $V(0) = 0$;
- (ii) $V(x) \rightarrow +\infty$ as $|x| \rightarrow \infty$;
- (iii) $\mathcal{L}_\phi V(x) \leq -|x|^2$, $x \neq 0$.

Call ϕ admissible if X_ϕ is stable. Just as in §3, we find that $\phi(x) = Kx$ is admissible if (3.8) has a positive solution P , and this is so whenever $G(x)$ is restricted by the inequality (3.14). Under these conditions the optimal linear control is determined exactly as before.

Appendix 1. Proof of Lemma 4.1. Let \mathcal{E}_x denote expectation on the paths of X when $x(0) = x \in E$. Let $t > 0$ be fixed and write

$$w(x) = \mathcal{E}_x\{v[x(t)]\}.$$

We show first that w exists a.e. $[\mu]$ and

$$(A1.1) \quad \mathcal{E}_\mu\{w\} = \mathcal{E}_\mu\{v\}.$$

If v is a simple function, (A1.1) is obvious. If $v \geq 0$ and v_n are simple functions with $0 \leq v_n \uparrow v$, then

$$w_n(x) = \mathcal{E}_x\{v_n[x(t)]\}$$

is measurable and $w_n \uparrow w$. By monotone convergence,

$$\begin{aligned} \mathcal{E}_\mu\{w\} &= \mathcal{E}_\mu\{\lim w_n\} = \lim \mathcal{E}_\mu\{w_n\} \\ &= \lim \mathcal{E}_\mu\{v_n\} = \mathcal{E}_\mu\{v\}. \end{aligned}$$

The general result follows by applying the argument to the positive and negative parts of v .

Now let v be of class $C^{(2)}$ and of compact support. By the Itô-Dynkin formula [4],

$$\begin{aligned} \mathcal{E}_\mu\left\{\mathcal{E}_x\left\{\int_0^t \mathcal{L}v[x(s)] ds\right\}\right\} &= \mathcal{E}_\mu\{\mathcal{E}_x\{v[x(t)] - v(x)\}\} \\ &= 0. \end{aligned}$$

Since $\mathcal{L}v[x(s)]$ is bounded and almost surely continuous (in s) there follows, by dominated convergence,

$$\begin{aligned} \mathcal{E}_\mu\{\mathcal{L}v(x)\} &= \mathcal{E}_\mu\left\{\mathcal{E}_x\left\{\lim_{t \downarrow 0} t^{-1} \int_0^t \mathcal{L}v[x(s)] ds\right\}\right\} \\ &= \lim_{t \downarrow 0} \mathcal{E}_\mu\left\{\mathcal{E}_x\left\{t^{-1} \int_0^t \mathcal{L}v[x(s)] ds\right\}\right\} \\ &= 0. \end{aligned}$$

In general, suppose $v(x)$ satisfies the integrability condition of the hypothesis. Then for any $\epsilon > 0$, there exists a smooth function $\tilde{v}(x)$ of compact support such that

$$|\varepsilon_\mu\{\mathcal{L}v(x)\} - \varepsilon_\mu\{\mathcal{L}\tilde{v}(x)\}| < \epsilon;$$

that is, $|\varepsilon_\mu\{\mathcal{L}v(x)\}| < \epsilon$.

Appendix 2. Proof of Theorem 6.1. It is assumed that X is a diffusion process determined by the stochastic differential equation (2.4); in particular, X satisfies the hypotheses made in [5] and [6]. For brevity we use freely the methods and notation of [6].

LEMMA. *Let $D \subset E$ be a normal domain with boundary Γ and let τ_Γ be the first time the path $x(\cdot)$ hits Γ . A necessary and sufficient condition that*

$$\varepsilon_x \left\{ \int_0^{\tau_\Gamma} L[x(t)] dt \right\} < \infty, \quad x \in E - D,$$

is that the equation

$$\mathcal{L}[u(x)] = -L(x), \quad x \in E - D,$$

have a smooth positive solution $u(x)$ in $E - D$.

Proof. The sufficiency was proved as Lemma 3.3 of [6]. To prove necessity let $\{D_n : n = 1, 2, \dots\}$ be a sequence of domains as in [6, Lemma 3.3], and put

$$u_n(x) = \varepsilon_x \left\{ \int_0^{\tau_n} L[x(t)] dt \right\}, \quad x \in D_n - \bar{D},$$

where $\tau_n = \min \{t : x(t) \in \Gamma \cup \Gamma_n, x(0) \in \bar{D}_n - D\}$. Since $\tau_n \uparrow \tau_\Gamma$, $n \rightarrow \infty$, with probability 1, there follows

$$u_n(x) \uparrow u(x) = \varepsilon_x \left\{ \int_0^{\tau_\Gamma} L[x(t)] dt \right\}, \quad x \in E - \bar{D}.$$

Since the $u_n(x)$ satisfy $\mathcal{L}u_n(x) = -L(x)$, $x \in D_n - \bar{D}$, $u_n(x) = 0$, $x \in \Gamma \cup \Gamma_n$, it follows (see [6, Proof of Lemma 3.1]) that $u(x) \geq 0$ has the required property.

Combining this result with that of [6, Lemma 3.2], we have that $\varepsilon_\mu\{L(x)\} < \infty$ if and only if there exists a smooth positive solution of

$$(A2.1) \quad \mathcal{L}u(x) = -L(x), \quad |x| > r,$$

for some $r < \infty$. The proof of Theorem 6.1 is completed by showing that no positive solution of (A2.1) exists, and follows almost verbatim the proof of [5, Theorem 4].

Acknowledgment. I am indebted to Dr. J. G. Heller for useful discussion of this problem.

REFERENCES

- [1] J. J. FLORENTIN, *Optimal control of continuous-time, Markov, stochastic systems*, J. Electronics Control, 10 (1961), pp. 473-488.

- [2] W. M. WONHAM, *Stochastic problems in optimal control*, Tech. Rep. 63-14, Research Institute for Advanced Studies, Baltimore, Maryland, 1963.
- [3] K. ITÔ, *On stochastic differential equations*, Mem. Amer. Math. Soc., no. 4, 1951, 51 pp.
- [4] E. B. DYNKIN, *Markov Processes*, Academic Press, New York, 1965.
- [5] W. M. WONHAM, *Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195-207.
- [6] ———, *A Liapunov method for the estimation of statistical averages*, Ibid., 2 (1966), pp. 365-377.
- [7] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.
- [8] R. BELLMAN, *Dynamic Programming*, Princeton University Press, Princeton, 1957.
- [9] R. A. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.
- [10] R. L. STRATONOVICH, *A new form of representation of stochastic integrals and equations*, this Journal, 4 (1966), pp. 362-371.
- [11] ———, *Conditional Markov Processes and their Applications in the Theory of Optimal Control*, Izd. Mosk. Univ., 1966.
- [12] E. WONG AND M. ZAKAI, *On the relation between ordinary and stochastic differential equations*, Internat. J. Engrg. Sci., 3 (1965), pp. 213-229.
- [13] R. Z. KHASHINSKII, *On the stability of the trajectory of Markov processes*, J. Appl. Math. Mech., 26 (1962), pp. 1554-1565.
- [14] C. E. LANGENHOP, *On the stabilization of linear systems*, Proc. Amer. Math. Soc., 15 (1964), pp. 735-742.

AN ITERATIVE PROCEDURE FOR SOLVING THE TIME-OPTIMAL REGULATOR PROBLEM*

TOSHIO FUJISAWA AND YUTAKA YASUDA†

1. Introduction. This paper is concerned with the computational aspects of the time-optimal regulator problem for linear dynamical systems with amplitude constraints. The nature of this problem has been theoretically made clear by LaSalle [1] and Pontryagin [2]. Several iterative procedures for computing the optimal control have been developed [3]–[8]. In this paper, a new solution method,¹ which is iterative and suitable for digital computers, is supplied with proof of convergence. The procedure does not depend on any subsidiary conditions, for instance, normality or regularity. The only restriction is that the system under consideration is controllable in a sense defined later. Moreover, exponential convergence of the main routine is assured under a fairly general condition.

Throughout this paper it is assumed that the motion of a linear dynamical system is described by the vector differential equation

$$(1.1) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

where x is the n -dimensional (column) state vector, u is the r -dimensional (column) control vector, $A(t)$ is a continuous $n \times n$ matrix, and $B(t)$ is a continuous $n \times r$ matrix. It is also assumed that every component of the control vector function $u(t)$ is a Lebesgue-measurable function of time t and the constraints to which the control vector is subject are of the form

$$(1.2) \quad |u^i(t)| \leq 1 \quad \text{for almost every } t, \quad i = 1, 2, \dots, r,$$

where $u^i(t)$ is the i th component of the control vector function $u(t)$. The set of all r -dimensional measurable vector functions which satisfy the constraints is denoted by U .

Denoting the fundamental $n \times n$ matrix of (1.1) by $X(t)$, and the product $X^{-1}(t)B(t)$ by $Y(t)$, the solution of (1.1) with initial condition $x(0) = x_0$ is given by

$$(1.3) \quad x(t, u) = X(t) \left[x_0 + \int_0^t Y(\tau)u(\tau) d\tau \right].$$

The time-optimal regulator problem to be considered in this paper may

* Received by the editors August 12, 1966, and in revised form April 17, 1967.

† Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, Japan. This work was supported in part by the Japan Ministry of Education under Grant-in-Aid for Institutional Research 91325.

¹ Reportedly Barr [9] obtained a similar iterative procedure independently of the present authors.

be stated as follows: Given an initial state x_0 , find the shortest possible time t^* for which $x(t^*, u^*) = 0$ (null vector) for some $u^* \in U$. Since the fundamental matrix $X(t)$ is nonsingular, the optimal time t^* is the shortest possible time for which the following equality holds for some $u \in U$:

$$(1.4) \quad -x_0 = \int_0^t Y(\tau)u(\tau) d\tau.$$

For every fixed $t \geq 0$, a subset $S(t)$ of the n -dimensional Euclidean space E_n is defined as follows:

$$(1.5) \quad S(t) = \left\{ x \mid x = \int_0^t Y(\tau)u(\tau)d\tau \text{ for some } u \in U \right\}.$$

The fundamental properties of $S(t)$ are summarized as follows [1].

THEOREM 1.² *The sets $S(t)$ satisfy the following properties:*

(P.1) $S(t)$ is a compact (bounded, closed) convex set in E_n ;

(P.2) $S(t) \supset S(t')$ whenever $t > t'$;

$$(P.3) \quad \max_{s \in S(t)} (\eta, s) = (\eta, \int_0^t Y(\tau) \overline{\text{sgn}} [Y^T(\tau)\eta] d\tau) \\ = \int_0^t |Y^T(\tau)\eta| d\tau;$$

(P.4) $S(t) = \overline{\cup_{0 \leq t' < t} S(t')}$, the closure of the set union of sets $S(t')$ for $0 \leq t' < t$.

A system is called controllable with respect to the initial state x_0 if there is at least one control satisfying the constraints (1.2) which brings the system from the initial state x_0 to the origin in some finite time. In the subsequent two sections the controllability with respect to the given initial state will be assumed. The following existence theorem has been given by LaSalle [1].

THEOREM 2. *The optimal control exists if the system is controllable, and then the optimal control is of the form*

$$(1.6) \quad u^*(t) = \overline{\text{sgn}} [Y^T(t)\eta^*], \quad 0 \leq t \leq t^*, \quad \eta^* \neq 0.$$

Geometrically speaking, the vector η^* in (1.6) is the outward normal vector of a supporting hyperplane of $S(t^*)$ passing through $-x_0$, which is a boundary point of $S(t^*)$.

2. An iterative procedure for finding an optimal control. According to the discussion in the preceding section, the time-optimal regulator problem

² For $\eta, s \in E_n$, (η, s) and $\|\eta\|$ denote the inner product and the norm in E_n , respectively, and $|\eta|$ denotes the special norm $\sum_{i=1}^n |\eta^i|$. Superscript T stands for transpose and the equality $\eta = \overline{\text{sgn}} s$ for $\eta^i = \overline{\text{sgn}} s^i$, $i = 1, 2, \dots, n$, where $\overline{\text{sgn}} s^i = 1$ if $s^i > 0$, -1 if $s^i < 0$, and $\overline{\text{sgn}} s^i$ is an arbitrary value in $[-1, 1]$ if $s^i = 0$.

By (2.2),

$$(2.4) \quad (a - v_k, v_k) = \max_{s \in S(t_k)} (a - v_k, s).$$

This means that the vector $a - v_k$, if it does not vanish, is the normal vector of the supporting hyperplane of $S(t_k)$ passing through the point v_k . Assuming $a - v_k \neq 0$, one obtains, by (2.4) and (P.3) of Theorem 1,

$$(2.5) \quad (a - v_k, a) > (a - v_k, v_k) = \int_0^{t_k} |Y^T(\tau)(a - v_k)| d\tau.$$

Since the system has been assumed to be controllable, there is a finite time $T > 0$ such that $a \in S(T)$. Therefore,

$$(2.6) \quad (a - v_k, a) \leq \max_{s \in S(T)} (a - v_k, s) = \int_0^T |Y^T(\tau)(a - v_k)| d\tau.$$

Based on the relations (2.5) and (2.6), one can easily conclude that a finite time $t_{k+1} (> t_k)$ is uniquely determined. Whenever $0 \leq t' < t_{k+1}$, $a \notin S(t')$ because of the minimality of t_{k+1} , and hence $t_{k+1} \leq t^*$.

The case of $a - v_k = 0$ and $a - v_{k-1} \neq 0$ is now considered. In this case, the value t_k has been determined to be the minimum of t_k for which the following equality holds:

$$(2.7) \quad (a - v_{k-1}, a) = \int_0^{t_k} |Y^T(\tau)(a - v_{k-1})| d\tau = \max_{s \in S(t_k)} (a - v_{k-1}, s).$$

Therefore, it is easy to see that $t_k = t^*$ and $a - v_{k-1} = \eta^*$.

Now the convergence of the procedure is demonstrated.

Proof of convergence. It suffices to prove $t_k \rightarrow t^*$ and $v_k \rightarrow a$, provided that the iteration does not terminate in a finite number of cycles. Let e_k be a unit vector defined by

$$(2.8) \quad e_k = \frac{(a - v_k)}{\|a - v_k\|}, \quad k = 0, 1, 2, \dots,$$

and let m_{k+1} be a point of $S(t_{k+1})$ for which

$$(2.9) \quad (e_k, m_{k+1}) = \max_{s \in S(t_{k+1})} (e_k, s).$$

Then from (2.3) and (2.4),

$$(2.10) \quad (e_k, m_{k+1}) = (e_k, a)$$

and

$$(2.11) \quad (e_k, v_k) = \max_{s \in S(t_k)} (e_k, s).$$

Using (2.10), (2.11) and the property (P.3), we obtain

$$\begin{aligned}
 (2.12) \quad \|a - v_k\| &= (e_k, a - v_k) = (e_k, a) - (e_k, v_k) \\
 &= (e_k, m_{k+1}) - (e_k, v_k) = \int_{t_k}^{t_{k+1}} |Y^T(\tau)e_k| d\tau.
 \end{aligned}$$

Hence,

$$(2.13) \quad \|a - v_k\| \leq (t_{k+1} - t_k) \max_{\substack{0 \leq \tau \leq t^* \\ \|e\|=1}} |Y^T(\tau)e|.$$

The maximum value on the right-hand side of (2.13) does exist, and the value is an absolute constant independent of k . Since the sequence $\{t_k\}$ is monotone increasing and bounded according to the controllability assumption, the difference $t_{k+1} - t_k$ converges to zero as $k \rightarrow \infty$, and hence, $\|a - v_k\|$ also converges to zero as $k \rightarrow \infty$. This implies the convergence of t_k to t^* .

It is shown below that any point of accumulation of the sequence $\{e_k\}$ can be used as η^* of inequality (2.1). Therefore, if there exists one and only one η^* for which (2.1) holds, then e_k converges to the value η^* .

Let $\{e_{k_n}\}$ be a converging subsequence of $\{e_k\}$ such that

$$(2.14) \quad \lim_{n \rightarrow \infty} e_{k_n} = e^*.$$

It suffices to show that

$$(2.15) \quad (e^*, a) \geq (e^*, s) \quad \text{for any } s \in S(t^*).$$

The negation of the statement (2.15) implies that there can be found a point $s^\infty \in S(t^*)$ for which

$$(2.16) \quad (e^*, s^\infty - a) = \epsilon > 0.$$

There exists a positive number N such that $n \geq N$ implies $\|e_{k_n} - e^*\| \leq \epsilon/(3\{\|s^\infty - a\| + \epsilon/3\})$. Then there exists a point $s' \in S(t_{k_{N'}})$, $N' \geq N$, such that $\|s^\infty - s'\| < \epsilon/3$, according to the property (P.4). Therefore,

$$\begin{aligned}
 (2.17) \quad (e_{k_{N'}}, s' - a) &= (e^*, s^\infty - a) \\
 &\quad + (e_{k_{N'}} - e^*, s' - a) + (e^*, s' - s^\infty).
 \end{aligned}$$

Applying Schwarz's inequality, (2.17) reduces to

$$(2.18) \quad (e_{k_{N'}}, s' - a) \geq \epsilon - \|e_{k_{N'}} - e^*\| \|s' - a\| - \|s' - s^\infty\| \geq \frac{\epsilon}{3} > 0,$$

which contradicts the relation $(e_{k_{N'}}, s' - a) \leq 0$.

So far the iterative procedure has been considered under the assumption of controllability. If the system is uncontrollable with respect to the given

initial state, that is, $a \notin \bigcup_{t \geq 0} S(t)$, the iterative procedure may fail at Step 2 by the fact that no finite value of t_{k+1} can give the equality (2.3). Then one can recognize that the system is uncontrollable with respect to the given initial state (see Fig. 2). The iterative procedure, however, may continue indefinitely without failures even in the uncontrollable case. In this case it can be shown that the sequence $\{t_k\}$ monotonically increases without bound and the sequence of points $\{v_k\}$ converges to a uniquely determined point a^* in the convex set $\overline{\bigcup_{t \geq 0} S(t)}$ for which $\|a - a^*\| \leq \|a - s\|$ holds for any $s \in \bigcup_{t \geq 0} S(t)$. Thus it seems to the authors that the iterative procedure itself has no ability to identify whether the system is controllable with respect to the given initial state.

The iterative procedures of [3]–[6] have been mainly concerned with the generation of the sequence of vectors η_k approximating the normal vector η^* to be found. The sequence of vectors $\{\eta_k\}$, in this paper, is given by the sequence $\{a - v_k\}$ as has been shown above. The method due to Eaton [4] and Pshenichniy [5] differs in Step 1 from the procedure described in this section. Using Eaton-Pshenichniy procedure, the modified vector η_{k+1} is any vector such that a hyperplane with normal vector η_{k+1} can strictly separate the point a and the set $S(t_k)$. The algorithm for this modification of normal vectors may be simpler compared with the procedure described in this paper, which necessitates the application of quadratic programming procedure to give a uniquely determined η_{k+1} . However, the underlying assumptions made in this paper are less restrictive: the system under consideration does not need to be normal or regular. Philosophically, the procedure in this paper is similar to the geometrical procedure given by Barr and Gilbert [10] for the minimum fuel problem.

3. Exponential convergence. It is theoretically interesting and also practically important to know the rate of the convergence of $\{t_k\}$ to t^* or of $\{\eta_k\}$ to η^* . Fadden and Gilbert [8] showed that an appropriate choice of

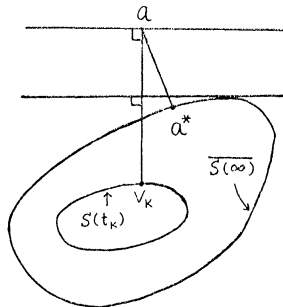


FIG. 2. Uncontrollable case ($\overline{S(\infty)}$ denotes $\overline{\bigcup_{t \geq 0} S(t)}$)

certain parameters leads to the exponential convergence of $\{\eta_k\}$ to η^* in Eaton-Pshenichnyi procedures [4], [5]. In this section, it is demonstrated under a fairly general condition that the convergence of $\{t_k\}$ to t^* is exponential using the procedure described in this paper.

Without loss of generality one can assume that the vector η^* is a unit vector, i.e.,

$$(3.1) \quad \|\eta^*\| = 1.$$

Therefore, applying Schwarz's inequality it follows that

$$(3.2) \quad (\eta^*, a - v_k) \leq \|a - v_k\|.$$

The following inequality is almost self-evident:

$$(3.3) \quad \begin{aligned} (\eta^*, a) - (\eta^*, v_k) &\geq (\eta^*, a) - \max_{s \in S(t_k)} (\eta^*, s) \\ &= \max_{s \in S(t^*)} (\eta^*, s) - \max_{s \in S(t_k)} (\eta^*, s). \end{aligned}$$

From the property (P.3) of Theorem 1, (3.2) and (3.3),

$$(3.4) \quad \|a - v_k\| \geq \int_{t_k}^{t^*} |Y^T(\tau)\eta^*| d\tau.$$

Now a sufficient condition is introduced to assure exponential convergence:

$$(3.5) \quad |Y^T(t^*)\eta^*| \neq 0.$$

Due to the assumption (3.5) and the relation (3.4) one can conclude there exists a positive number $\epsilon > 0$ such that $t^* - t_k < \epsilon$ implies

$$(3.6) \quad \|a - v_k\| \geq \frac{1}{2}(t^* - t_k) |Y^T(t^*)\eta^*|.$$

Combining this with (2.13), we obtain

$$(3.7) \quad \frac{t_{k+1} - t_k}{t^* - t_k} \geq \frac{|Y^T(t^*)\eta^*|}{2 \max_{\substack{0 \leq \tau \leq t^* \\ \|\epsilon\|=1}} |Y^T(\tau)\eta^*|}.$$

Denoting the right-hand side of the above inequality by θ it is clear that $0 < \theta < 1$. From this it follows that

$$(3.8) \quad \frac{t^* - t_{k+1}}{t^* - t_k} = 1 - \frac{t_{k+1} - t_k}{t^* - t_k} \leq 1 - \theta,$$

i.e., $t^* - t_{k+1} \leq (1 - \theta)(t^* - t_k)$. This conclusion assures that the sequence $\{t_k\}$ converges exponentially to t^* in a neighborhood of t^* .

The sufficient condition (3.5) is fairly general, and one can expect to

have exponential convergence of $\{t_k\}$ to t^* almost always when applying the iterative procedure proposed in this paper.

Dem'yanov [7] used an approach quite different from others [3]–[6]. Using his approach, one always has to keep a pair of time instants (t_k^l, t_k^u) , where $t_k^l < t^* < t_k^u$ and the error size $t_k^u - t_k^l$ is cut in half after each iteration. Therefore, the convergence of $\{t_k^l\}_{k=1,2,\dots}$ to the optimal time t^* is exponential. Difficulties may arise in trying to obtain an initial guess about time t_1^u .

4. Numerical computation and an illustrative example. Theoretically, an infinite number of iterations are necessary for determining the point v_k in Step 1 of the procedure in §2 by the application of the quadratic programming method [11]–[14]. Thus, there must be available a theoretical and practical rule for deciding when one has to terminate the iteration and, furthermore, there must exist a theoretical assurance of the convergence of the modified procedure with this rule. The rule adopted here is identical with the one given by Barr and Gilbert [10] for the minimum fuel problem. For the purpose of verifying the validity of this rule for the procedure of this paper, it is required to look into some details of the quadratic programming method [11]–[14].

The method can be started with an arbitrary point $v_{k_0} \in S(t_k)$ with associated control u_{k_0} . At the n th cycle, the point $v_{k_n} \in S(t_k)$ is known with associated control u_{k_n} . The one-cycle of the iteration consists of the following two steps.

Step Q-1. Find a point $w_{k_n} \in S(t_k)$ with associated control z_{k_n} such that

$$(4.1) \quad (a - v_{k_n}, w_{k_n}) = \max_{s \in S(t_k)} (a - v_{k_n}, s).$$

If $(a - v_{k_n}, w_{k_n} - v_{k_n}) = 0$, then let $v_{k_n} = v_k$, and the iteration terminates. If otherwise, go to the following step.

Step Q-2. Find θ_n , $0 \leq \theta_n \leq 1$, such that

$$(4.2) \quad \|a - \{(1 - \theta_n)v_{k_n} + \theta_n w_{k_n}\}\| = \min_{0 \leq \theta \leq 1} \|a - \{(1 - \theta)v_{k_n} + \theta w_{k_n}\}\|.$$

Then let $v_{k_{n+1}} = (1 - \theta_n)v_{k_n} + \theta_n w_{k_n}$ and $u_{k_{n+1}} = (1 - \theta_n)u_{k_n} + \theta_n z_{k_n}$, and go back to Step Q-1.

From (4.1) and (4.2), the following relations may be easily derived [11]–[14]:

$$(4.3) \quad \|v_k - v_{k_n}\|^2 \leq \|a - v_{k_n}\|^2 - \|a - v_k\|^2 \leq 2(a - v_{k_n}, w_{k_n} - v_{k_n}),$$

$$(4.4) \quad (a - v_{k_n}, w_{k_n} - v_{k_n}) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Under the assumption of $\|a - v_k\| \neq 0$, and for a positive number μ less

than unity which is fixed in advance, there exists a positive integer N such that $n \geq N$ implies

$$(4.5) \quad (a - v_{k_n}, w_{k_n} - v_{k_n}) \leq \mu \|a - v_k\|^2.$$

Since $\|a - v_k\| \leq \|a - v_{k_n}\|$ for any n , it follows that for $n \geq N$,

$$(4.6) \quad (a - v_{k_n}, w_{k_n} - v_{k_n}) \leq \mu \|a - v_{k_n}\|^2.$$

Any point of the sequence $\{v_{k_n}\}_{n=1,2,\dots}$, which satisfies (4.6), can serve as a good approximation to the point v_k . This is the rule mentioned above. Using (4.6), we obtain

$$\begin{aligned} (4.7) \quad & (a - v_{k_n}, a) - \max_{s \in S(t_k)} (a - v_{k_n}, s) \\ &= (a - v_{k_n}, a) - (a - v_{k_n}, w_{k_n}) \\ &= (a - v_{k_n}, a - v_{k_n}) - (a - v_{k_n}, w_{k_n} - v_{k_n}) \\ &\geq \|a - v_{k_n}\|^2 - \mu \|a - v_{k_n}\|^2 = (1 - \mu) \|a - v_{k_n}\|^2 > 0, \end{aligned}$$

which assures the application of Step 2 of the main iteration to be possible. Then, it is not difficult to reconstruct the proof of the convergence of the modified procedure with the use of the approximations v_{k_n} instead of the exact solutions v_k , provided that $a \notin S(t_k)$ for any $k = 1, 2, \dots$. The property of exponential convergence of the main iteration also remains valid. The outline of the reconstruction is as follows. Inequality (3.4) holds with v_k replaced by v_{k_n} , and due to (4.6), one can derive for the substitution of (2.12),

$$(4.8) \quad (1 - \mu) \|a - v_k\| \leq \int_{t_k}^{t_{k+1}^*} |Y^T(\tau) e'_{k_n}| d\tau,$$

where $e'_{k_n} = (a - v_{k_n}) / \|a - v_{k_n}\|$. Using these relations, the proof of §§2 and 3 can easily be reconstructed.

If $a \in S(t_k)$ at the k th cycle of the main iteration, then $t^* = t_k$ and $a = v_k$. In this situation, there are two possible cases to be considered. The first is the case where the minor iteration terminates with v_{k_n} satisfying (4.6). Then, Step 2 of the main iteration is applied, and this results in the answer $t_{k+1} = t_k$, which tells $t^* = t_k$. The second is the case where the minor iteration continues indefinitely because no v_{k_n} satisfying (4.6) can be found, and then $v_{k_n} \rightarrow v_k = a$ as $n \rightarrow \infty$. Regardless of which case actually occurs, the optimal time $t^* = t_k$ is obtained and the control associated with v_{k_n} may be considered as a practical answer. This completes the proof of the validity of the adopted rule.

There is another important property which is of value for numerical computation. The property is that the exact determination of t_{k+1} is not

necessary. One may choose any value t'_{k+1} such that $t_k < t'_{k+1} \leq t_{k+1}$ and

$$(4.9) \quad \int_0^{t'_{k+1}} |Y^T(\tau)(a - v_{k_n})| d\tau \geq (a - v_{k_n}, \alpha a + (1 - \alpha)v_{k_n})$$

for $k = 1, 2, \dots$,

where α is a fixed number lying between μ and unity ($\mu < \alpha \leq 1$) and v_{k_n} is a good approximation to v_k . The nature of exponential convergence is not violated. The proof of this, which is not described here, is not so difficult. The facts mentioned above can be successfully utilized to cut down the machine time considerably.

As an illustrative example the system characterized by

$$(4.10) \quad \dot{x} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} 1 \\ 1 \end{pmatrix} u$$

is considered. Matrix (AB, B) is nonsingular and hence system (4.10) is controllable in the ordinary sense. Also $Y(\tau) = e^{-A\tau}B = \begin{pmatrix} 1 \\ e^{-\tau} \end{pmatrix}$ and $Y^T(\tau)\eta = \eta^1 + \eta^2 e^{-\tau}$, and thus the system is normal. Moreover, for this system analytical determination of $S(t)$ and the solution of maximizing (η, s) on $S(t)$ may be easily derived.

The computational results are shown in Tables 1, 2 and Fig. 3. In (i) of Table 1 and Fig. 3 are shown results of iterations for a controllable case with respect to the initial state $-a$, where after four iterations the error $\|a - v_k\|$ was less than the preassigned 10^{-3} . For this example, the result of iteration of the procedure for minimizing $\|a - s\|$ on $S(t)$ is shown in Table 2 and Fig. 3. In (ii) and (iii) of Table 1 are shown the results for uncontrollable cases. The iteration terminated in five or three cycles, satisfying $t_{k+1} - t_k \leq 0.0001$ in the case (ii) or $t \geq 20$ in the case (iii), respectively.

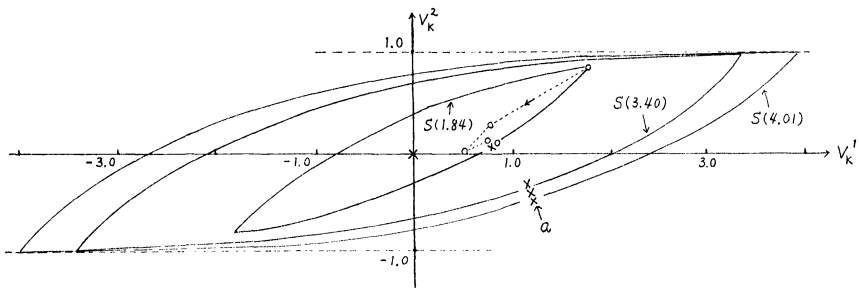


FIG. 3. Results of iteration, $a = (1.2, -0.55)$. (The mark \times denotes v_k and the mark \circ denotes $v_{k_n} \in S(1.84)$.)

TABLE 1
Results of iteration
(i) $a = (1.2, -0.55)$

Number of Cycles	Time t_k	State $v_k = (v_k^1, v_k^2)$		$\ a - v_k\ $
		v_k^1	v_k^2	
0	0.00000	0.00000	0.00000	1.32004
1†	1.83744	0.83556	0.05272	0.70433
2	3.40427	1.14715	-0.38623	0.17208
3	4.01138	1.19481	-0.52897	0.02166
4‡	4.10831	1.19978	-0.54911	0.00091

(ii) $a = (1.2, -1.0)$

0	0.00000	0.00000	0.00000	1.56205
1	2.81684	0.92211	-0.28427	0.76778
2	5.07610	1.16097	-0.72384	0.27891
3	7.10733	1.19463	-0.89679	0.10335
4	9.11136	1.20120	-0.92753	0.07248
5	9.11136	1.20120	-0.92753	0.07248

(iii) $a = (1.2, 1.5)$

0	0.00000	0.00000	0.00000	1.92094
1	1.99501	1.33977	0.75854	0.75453
2	6.78357	1.20980	0.96438	0.53571
3	$t_k \geq 20$			

† See Table 2.

‡ $\eta^* = (22, -89)$ and hence $u^*(\tau) = -1$ for $0 \leq \tau \leq \tau_0$ and 1 for $\tau > \tau_0$, where $\tau_0 = 1.398$.

TABLE 2
Results of iteration for minimizing $\|a - s\|$ on $S(1.83744)$

Number of Cycles	State $v_{k_n} = (v_{k_n}^1, v_{k_n}^2)$		$\ a - v_{k_n}\ $	$(a - v_{k_n}, w_{k_n} - v_{k_n})$
	$v_{k_n}^1$	$v_{k_n}^2$		
0	1.83744	0.84078	1.52990	4.68117
1	0.78415	0.35881	0.99944	0.33568
2	0.50957	0.03443	0.90457	0.44555
3	0.75471	0.18329	0.85790	0.13177
4	0.83980	0.05527	0.70434	0.00015
5	0.83711	0.05365	0.70433	0.00004
6	0.83556	0.05272	0.70433	0.00001

Acknowledgment. The authors wish to thank the reviewers for their helpful comments. Much of the material in §4 was added after the first review on the suggestion of one of the reviewers.

REFERENCES

- [1] J. P. LaSALLE, *The time-optimal control problem*, Contributions to the Theory of Nonlinear Oscillations, vol. 5, Princeton University Press, Princeton, 1960, pp. 1-24.
- [2] L. S. PONTYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
- [3] L. W. NEUSTADT, *Synthesis of time-optimal control systems*, J. Math. Anal. Appl., 1 (1960), pp. 484-492.
- [4] J. H. EATON, *An iterative solution to time-optimal control*, Ibid., 5 (1962), pp. 329-344.
- [5] B. N. PSHENICHNIY, *A numerical method of computing time-optimal controls in linear systems*, Zh. Vychisl. Mat. i Mat. Fiz., 4 (1964), pp. 52-60.
- [6] T. G. BABUNASHVILI, *The synthesis of linear optimal systems*, this Journal, 2 (1964), pp. 261-265.
- [7] V. F. DEM'YANOV, *Determination of the optimum program in a linear system*, Automat. i Telemekh., 25 (1964), pp. 3-11.
- [8] E. J. FADDEN AND E. G. GILBERT, *Computational aspects of the time-optimal control problem*, Computing Methods in Optimization Problems, A. V. Balakrishnan and L. W. Neustadt, eds., Academic Press, New York, 1964, pp. 167-192.
- [9] R. O. BARR, *Computation of optimal controls on convex reachable sets*, Proc. Conference on Mathematical Theory of Control, University of Southern California, 1967, Academic Press, New York, to appear.
- [10] R. O. BARR AND E. G. GILBERT, *Some iterative procedures for computing optimal controls*, Proc. Third Congress of the International Federation of Automatic Control, London, 1966.
- [11] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1956), pp. 95-110.
- [12] C. BERGE AND A. GHOUILA-HOURI, *Programming, Games and Transportation Networks*, John Wiley, New York, 1965.
- [13] P. S. FANCHER, *Iterative computation procedures for an optimum control problem*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 346-348.
- [14] E. G. GILBERT, *An iterative procedure for computing the minimum of quadratic form on a convex set*, this Journal, 4 (1966), pp. 61-80.

A NOTE ON THE FIXED-POINT METHOD OF J. E. RUBIO*

ELMER G. GILBERT†

In a recent paper J. E. Rubio [1] outlines a computational procedure for minimizing the Euclidean norm of the terminal state of a linear time-varying system. The purpose of this note is to point out an error in the proof of the key theorem (Theorem 3) and to show by means of a simple example that under very reasonable conditions (stronger than the hypotheses of the paper) the proposed computational procedure will not converge.

Wherever possible the subsequent notation is identical to the notation used in [1].

Consider first the error in the proof of Theorem 3. For the conclusion of part (i) to be valid it is necessary to show that there exists a fixed $\rho > 0$ such that for any $\bar{\epsilon} > 0$ there exists an $\bar{N}(\bar{\epsilon})$ such that

$$\|L_{\sigma_2}^p c - L_{\sigma_2}^q c\| < \bar{\epsilon}$$

for all $|\sigma_1 - \sigma_2| < \rho$ and $p, q > \bar{N}(\bar{\epsilon})$. The arguments of part (i) have not accomplished this since $|\sigma_1 - \sigma_2| < \min(\delta_1(\epsilon_1), \delta_2(\epsilon_2))$ and ϵ_1 and ϵ_2 are dependent on the choice of $\bar{\epsilon}$.

The following example shows that it is impossible to prove convergence of $\{L_0^m c\}$ under the hypotheses of the paper. Take $n = 2, r = 1$ and

$$A(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B(t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad t_0 = 0, \quad t_f = 2\pi.$$

Then a simple calculation shows that

$$(1) \quad x_1(t_f, u_c^*) = x_{10} - 4c_1, \quad x_2(t_f, u_c^*) = x_{20} - 4c_2, \quad c \in \Omega,$$

which is a parametric representation of a circle with radius 4 and center x_0 . To satisfy the assumption that $\|x(t_f, u^*)\| > 0$ and to simplify the following developments, choose $x_{10} = \rho > 4$ and $x_{20} = 0$. If points $c \in \Omega$ are denoted by $c_1 = \cos \theta, c_2 = \sin \theta, \theta \in (-\pi, \pi]$, it is easy to show that $L_0 c$ corresponds to the function

$$(2) \quad F(\theta) = -\tan^{-1} [4(\rho - 4 \cos \theta)^{-1} \sin \theta], \quad \theta \in (-\pi, \pi],$$

where it is understood that $\tan^{-1}(\cdot)$ is into $(-\frac{1}{2}\pi, \frac{1}{2}\pi)$. Thus the sequence $\{L_0^m c\}$ corresponds to the solution of the difference equation

* Received by the editors February 9, 1967, and in revised form April 28, 1967.

† Information and Control Engineering, University of Michigan, Ann Arbor, Michigan. This work was sponsored by the United States Air Force under Grant AF-AFOSR-814-66.

$$(3) \quad \theta_{k+1} = F(\theta_k), \quad \theta_0 \in (-\pi, \pi]$$

with $c_1 = \cos \theta_0$ and $c_2 = \sin \theta_0$.

The solution of (3) exhibits two qualitatively different behaviors. For $\rho \geq 8$ and $\theta_0 \in (-\pi, \pi]$ or $4 < \rho < 8$ and $\theta_0 = 0$, $\theta_k \rightarrow 0$, which is the desired result since (1) shows that $x(t_f, u_c^*) = x(t_f, u^*)$ for $c_1 = 1$, $c_2 = 0$. For $4 < \rho < 8$, $\theta_0 \in (-\pi, \pi]$, $\theta_0 \neq 0$, θ_k tends to one of the following two sequences: $\{(-1)^m \bar{\theta}\}$ or $\{-(-1)^m \bar{\theta}\}$, where $\bar{\theta} \in (0, \pi/2)$ is the (unique) solution of $\theta = -F(\theta)$ on $(0, \pi/2)$. Thus the fixed-point method fails for $4 < \rho < 8$ unless $\theta_0 = 0$. The above results are deduced from (3) by using the following easily confirmed facts:

- (i) $\theta_k \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$, $k \geq 1$;
- (ii) $F(-\theta) = -F(\theta)$;
- (iii) $(dF/d\theta)|_{\theta=0} = 4(4 - \rho)^{-1}$;
- (iv) $dF/d\theta$ is a strictly increasing function on $[0, \frac{1}{2}\pi)$;
- (v) $0 < F(-\pi/2) < 1$.

The fixed-point method is very similar to the method described by Gilbert [2]. From the assumption $\min_{u \in U} \|x(t_f, u)\| \neq 0$ it follows that for all $c \neq 0$, $x(t_f, u_c^*) \neq 0$. This plus the homogeneity of $S_c(u)$ in c implies that the operator Q , which takes c into $x(t_f, u_c^*)$, generates for $c \neq 0$ a sequence $\{Q^m c\}$ whose elements equal the corresponding elements of $\{L_0^m c\}$ multiplied by appropriate positive scalars. But the elements $Q^m c$ are those produced by [2, (3.3)] if $z_0 = c$ and $\alpha_m = 1$, $m = k \geq 0$. In most applications (3.4) of [2] requires $0 < \alpha_m < 1$ for some m . In these cases the procedure of [2] will converge more rapidly in the sense that $\|x(t_f, u_{Q^m c})\| < \|x(t_f, u_{L_0^m c})\|$. Thus, even if conditions can be found under which the fixed-point method will converge, it will be less efficient than the procedure of [2]. Moreover, the procedure of [2] is applicable to a much wider class of problems.

REFERENCES

- [1] J. E. RUBIO, *A fixed-point method for a minimum-norm control problem*, this Journal, 4 (1966), pp. 705-715.
- [2] E. G. GILBERT, *An iterative procedure for computing the minimum of a quadratic form on a convex set*, this Journal, 4 (1966), pp. 61-80.

THE EXISTENCE OF PIECEWISE CONTINUOUS FUEL OPTIMAL CONTROLS*

W. C. GRIMMELL†

Introduction. From an engineering point of view knowing that there exists a measurable function which is a solution of an optimal control problem is almost as unsatisfactory as having no knowledge concerning the question of existence. The engineer should be interested in the existence of continuous or at least piecewise continuous optimal controls. Unfortunately, most proofs of the existence of optimal controls are proofs of the existence of measurable optimal controls. An exception is a proof by Halkin [1] of the existence of "bang-bang" piecewise continuous time optimal controls for the class of linear systems in which the elements of the coefficient matrices are piecewise analytic functions. Using Halkin's approach, the present paper demonstrates the existence of piecewise continuous finitely valued fuel optimal controls for fixed time fuel optimal problems involving the above systems.

Problem description. We consider a system described by a matrix differential equation of the form

$$(1) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t),$$

where $x(t)$ is an n -dimensional state vector, $u(t)$ is an m -dimensional control vector, and A and B are matrices of piecewise analytic functions defined on the interval $[0, T]$.

A function f on $[0, T]$ is piecewise analytic if there are a finite set $\{t_0, t_1, \dots, t_k\}$ with $0 = t_0 < t_1 < t_2 < \dots < t_k = T$, a finite collection of functions $f^1(t), f^2(t), \dots, f^k(t)$, and a $\delta > 0$ such that

- (i) $f(t) = f^i(t)$ for all $t \in (t_{i-1}, t_i)$ and each $i = 1, 2, \dots, k$;
- (ii) $f^i(t)$ is defined and analytic on $(t_{i-1} - \delta, t_i + \delta)$ for each $i = 1, 2, \dots, k$.

Admissible controls for the system are measurable functions from $[0, T]$ to R^m whose values lie in $\Omega = \{\alpha \in R^m : |\alpha_i| \leq 1, i = 1, 2, \dots, m\}$. Admissible controls whose instantaneous values lie in $\Gamma = \{\alpha \in R^m : \alpha_i \in \{-1, 0, 1\}, i = 1, 2, \dots, m\}$ will be called zero-extreme controls.

A fixed time fuel optimal control problem calls for the finding of a control u which transfers the system from a given initial state $x(0)$ to a

* Received by the editors March 29, 1967.

† Underwater Systems Analysis Department, Bell Telephone Laboratories, Whippany, New Jersey 07981.

given final state $x(T)$ and minimizes the cost functional¹

$$J(u) = \int_0^T \left(\sum_{i=1}^m |u_i(t)| \right) dt$$

over the set of admissible controls causing the specified transfer.

We will show that for the fixed time fuel optimal problem any state transfer which can be accomplished with an admissible control can be accomplished optimally with a zero-extreme piecewise continuous control.

Preliminaries. The proof in the next section is accomplished by viewing the minimum fuel problem with its nonlinear cost functional in terms of appropriate linear problems. The following lemma proved in [1] is then applicable.

LEMMA. *Let M be the set of all real-valued measurable functions y on $[0, T]$ with $0 \leq y(t) \leq 1$. Let M' be the subset of piecewise continuous functions in M with $|y(t) - \frac{1}{2}| = \frac{1}{2}$. Let ψ be an r -dimensional piecewise analytic vector function on $[0, T]$. Define*

$$K = \left\{ \int_0^T \psi(t)y(t) dt : y \in M \right\},$$

$$K' = \left\{ \int_0^T \psi(t)y(t) dt : y \in M' \right\}.$$

Then $K = K'$.

Theoretical discussion. For any state transfer, the existence of an admissible control which causes the transfer implies the existence of a fuel optimal (measurable) control for the transfer [3].² Let this optimal control be denoted by v . Let $\phi^i(t)$ denote the i th column of $\Phi^{-1}(t, 0)B(t)$, where $\Phi(t, 0)$ is the state transition matrix of (1),³ and let

$$\xi^i = \int_0^T \phi^i(t)v_i(t) dt.$$

¹ In [2] it was shown that a control which is optimal with respect to the cost functional specified above is also optimal with respect to any cost functional in a general class of functionals. The results of the present paper are therefore applicable to a much wider class of problems than that described above.

² From [3] the existence of an optimal control follows immediately. An additional result may be obtained. Let F be the function from R^n to R^{n+1} which is defined as follows: if $\alpha = F(\beta)$, then $\alpha_i = \beta_i$, $i = 1, 2, \dots, n$, and $\alpha_{n+1} = \sum_{i=1}^n |\beta_i|$. The existence of a zero extreme (measurable) control which is optimal follows from [3, p. 115] once it is noted that the sets $\{\alpha \in R^{n+1} : \alpha = F(\beta), \beta \in \Omega\}$ and $\{\alpha \in R^{n+1} : \alpha = F(\beta), \beta \in \Gamma\}$ have the same closed convex hull.

³ The value of $x(t)$ is given by

$$x(t) = \Phi(t, 0) \left(x(0) + \int_0^t \Phi^{-1}(s, 0)B(s)u(s) ds \right).$$

Note that ϕ^i is an n -dimensional piecewise analytic vector function.

A measurable scalar-valued function, defined on $[0, T]$, whose instantaneous values lie in $[-1, 1]$ will be called an admissible component.

THEOREM. *There exists, for each $i, i = 1, 2, \dots, m$, a piecewise continuous admissible component w_i such that*

$$\int_0^T \phi^i(t) w_i(t) dt = \zeta^i, \quad \int_0^T |w_i(t)| dt = \int_0^T |v_i(t)| dt,$$

and for every $t \in [0, T]$, $w_i(t) \in \{-1, 0, 1\}$.

Proof. Let $G = \{t \in [0, T]: v_i(t) > 0\}$, $H = \{t \in [0, T]: v_i(t) < 0\}$ and $N = \{t \in [0, T]: v_i(t) = 0\}$. Note $G \cup H \cup N = [0, T]$. Let

$$v_i^p(t) = \begin{cases} v_i(t) & \text{for } t \in G, \\ 0 & \text{for } t \in (H \cup N), \end{cases}$$

$$v_i^m(t) = \begin{cases} v_i(t) & \text{for } t \in H, \\ 0 & \text{for } t \in (G \cup N). \end{cases}$$

Also let

$$\int_0^T \phi^i(t) v_i^p(t) dt = \zeta^{ip} \quad \text{and} \quad \int_0^T \phi^i(t) v_i^m(t) dt = \zeta^{im}.$$

Note that $v_i(t) = v_i^p(t) + v_i^m(t)$ and hence, $\zeta^i = \zeta^{ip} + \zeta^{im}$. Furthermore,

$$\int_0^T |v_i(t)| dt = \int_0^T (v_i^p(t) - v_i^m(t)) dt.$$

Let $\theta^i(t)$ be an $(n+1)$ -dimensional vector with $\theta_j^i(t) = \phi_j^i(t)$, $j = 1, 2, \dots, n$, and $\theta_{n+1}^i(t) = 1$. Then for any admissible component u_i ,

$$\int_0^T \theta^i(t) u_i(t) dt = \int_0^T \begin{bmatrix} \phi^i(t) \\ 1 \end{bmatrix} u_i(t) dt = \begin{bmatrix} \alpha^i \\ a_i \end{bmatrix},$$

where α^i is an n -dimensional vector and a_i is a scalar. That is,

$$\alpha^i = \int_0^T \phi^i(t) u_i(t) dt \quad \text{and} \quad a_i = \int_0^T u_i(t) dt.$$

Now θ^i is an $(n+1)$ -dimensional piecewise analytic vector function and hence, by the lemma, there exists a piecewise continuous component w_i^p , whose values are 0 and 1, and a piecewise continuous component w_i^m , whose values are 0 and -1 , such that

$$\int_0^T \theta^i(t) w_i^p(t) dt = \int_0^T \theta^i(t) v_i^p(t) dt$$

and

$$\int_0^T \theta^i(t) w_i^m(t) dt = \int_0^T \theta^i(t) v_i^m(t) dt.$$

From the above we see that these equalities are equivalent to the equalities

$$\begin{aligned} \int_0^T \phi^i(t) w_i^p(t) dt &= \zeta^{ip}, & \int_0^T w_i^p(t) dt &= \int_0^T v_i^p(t) dt, \\ \int_0^T \phi^i(t) w_i^m(t) dt &= \zeta^{im}, & \int_0^T w_i^m(t) dt &= \int_0^T v_i^m(t) dt. \end{aligned}$$

Let $w_i(t) = w_i^p(t) + w_i^m(t)$. Note that w_i is a piecewise continuous component whose values are 0, 1 and -1 . Also,

$$\int_0^T \phi^i(t) w_i(t) dt = \zeta^{ip} + \zeta^{im} = \zeta^i$$

and

$$\int_0^T |w_i(t)| dt \leq \int_0^T (w_i^p(t) - w_i^m(t)) dt = \int_0^T |v_i(t)| dt.$$

But since v is optimal,

$$\int_0^T |w_i(t)| dt \geq \int_0^T |v_i(t)| dt.$$

Hence,

$$\int_0^T |w_i(t)| dt = \int_0^T |v_i(t)| dt.$$

This completes the proof of the theorem.

The control w whose components are the w_i of the theorem will cause the same transfer as the control v . Also, $J(w) = J(v)$. Hence, w is a piecewise continuous zero-extreme control which is optimal.

Conclusion. In summary, this paper demonstrates that for the fuel optimal problem any state transfer which can be accomplished with an arbitrary admissible control can be accomplished optimally with a piecewise continuous zero-extreme control. In terms of engineering design this corresponds to a three-position contactor control with a finite number of switchings.

The results of the paper are quite general, applying even to cases in which the possibility of singular fuel optimal controls exists (see [4, pp. 437, 481–496]). For a system having a form which excludes the possibility of singular controls, application of the maximum principle leads directly to the conclusion that all fuel optimal controls are zero-extreme piecewise continuous functions. However, in cases in which singular controls exist, the maximum

principle does not guarantee that, for any achievable state transfer, at least one fuel optimal control is a zero-extreme piecewise continuous function. Thus, in demonstrating this, the present paper eliminates a portion of the ambiguity resulting from the direct application of the maximum principle.

REFERENCES

- [1] H. HALKIN, *A generalization of LaSalle's bang-bang principle*, this Journal, 2 (1965), pp. 199-202.
- [2] W. C. GRIMMELL AND W. L. NELSON, *Optimal control problems having identical solutions*, IEEE Trans. Automatic Control, AC-11 (1966), pp. 751-752.
- [3] L. W. NEUSTADT, *The existence of optimal controls in the absence of convexity conditions*, Math. Anal. Appl., 7 (1963), pp. 110-117.
- [4] M. ATHANS AND P. L. FALB, *Optimal Control: An Introduction to the Theory and Its Applications*, McGraw-Hill, New York, 1966.

OPTIMAL DISCOUNTED STOCHASTIC CONTROL FOR DIFFUSION PROCESSES*

H. J. KUSHNER†

Summary. We consider the problem of controlling the random process defined by (1) with the so-called "discounted" cost criterion (3). Conditions are given under which an optimal Lipschitz continuous control exists. There is an associated partial differential equation whose "solution" is the "optimal cost". However, the problem is not well posed since the domain on which the solution is to be defined (R^n) is unbounded, and no boundary conditions are given. However, one can define a sequence of well-posed Dirichlet problems, in a sequence of sets $S_i \rightarrow R^n (S_i \supset S_{i-1})$, so that the limit of the sequence (which is the optimal cost) yields the optimal control.

Under weaker conditions than required for the above results, a condition is given which is sufficient to assure the optimality of a given "cost" and control. The criterion provides a relatively simple test by which a formally obtained control and cost may be rigorously checked for optimality. Finally, the solution to the discounted problem for a linear-quadratic problem is obtained in a straightforward manner.

The surveys [12], [13], [14] contain some background material on the control of diffusion processes and numerous references.

1. Introduction and assumptions. Let

$$(1) \quad dx = f(x, u) dt + \sigma(x) dz$$

be an (Itô) stochastic differential equation with differential generator:

$$(2) \quad \begin{aligned} L^u &= \sum_i f_i \frac{\partial}{\partial x_i} + \sum_{i,j} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j}, \\ a_{ij} &= \frac{1}{2} \sum \sigma_{i\mu} \sigma_{\mu j}, \end{aligned}$$

and with the $z_i(t)$ independent Weiner processes and f and σ satisfying uniform Lipschitz conditions in x and u (for f only). Define $S_n = \{x: |x| \leq n\}$, where $|x|^2 = x'x$. Controls $u = u(x)$ with values in the compact convex set U and satisfying a local Lipschitz condition in x are called *admissible*. Note that the escape time for the x_t process corresponding to any admissible control is infinite with probability one (w. p. 1).¹

* Received by the editors March 1, 1967, and in revised form May 1, 1967.

† Division of Applied Mathematics, Center for Dynamical Systems, Brown University, Providence, Rhode Island 02912. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grant AF-AFOSR-693-66, in part by the National Science Foundation, Engineering Branch, under Grant GK-967, and in part by the National Aeronautics and Space Administration under Grant NGR-40-002-015.

¹ Escape time is the *time of escape of x_t to ∞* ; i.e., let $\tau_n = \inf\{t: |x_t| \geq n\}$, with $\tau_n = \infty$ if $\sup |x_t| < n$. Then τ_n is nondecreasing w.p. 1 and the escape time is $\lim \tau_n$ (defined w.p.1). Essentially, the Lipschitz conditions together with the boundedness of U imply that $E \max_{t \leq T} x_t' x_t \leq K e^{CT}$ for some real positive K and C , and this implies that the escape time is infinite.

Often x_t , u_t and u are written for $x(t)$, $u(x(t))$ and $u(x)$, respectively. The cost associated with each u is the "discounted" cost function

$$(3) \quad C^u(x) = E_x^u \int_0^\infty e^{-\beta t} k(x_t, u_t) dt, \quad \beta > 0, \quad k \geq 0.$$

E_x^u is the expectation for the process corresponding to control u and starting in state $x = x_0$. $k(x, u)$ is supposed bounded in any compact x set, for any admissible u , and is locally Lipschitz in x and u . Furthermore, we suppose that there is some admissible u for which $C^u(x) < \infty$ for all x .²

In §3, we require that L^u be elliptic; i.e., in each compact x set there is a constant $\mu > 0$ so that

$$(4) \quad \sum a_{ij} \xi_i \xi_j \geq \mu |\xi|^2$$

for each vector ξ . This is dropped in §4.

Admissibility is defined as it is because the (Itô) process corresponding to discontinuous $u(x)$ has not yet been defined. To assure admissibility of our derived controls, some rather strong conditions seem to be required, in particular, (5) and (6):

$$(5) \quad f(x, u) \text{ is linear in } u;$$

$$(6) \quad \sum_{i,j} \left(\frac{\partial^2 k(x, u)}{\partial u_i \partial u_j} \right) \mu_i \mu_j \geq c |\mu|^2$$

for some $c > 0$ and all controls μ , and uniformly in x and u .

There are examples where the presumed "optimum" u is discontinuous when (5) and (6) are violated; e.g., the scalar problem $f = x + u$, $k = x^2 + |u|$. There is very likely a way of modeling and treating the stochastic control problem when the control is discontinuous, but, here, in order to stay within the domain where (1) is defined, we are obliged to require (5) and (6). Note that the dependence on x , of both f and k , is virtually unrestricted.

2. Outline. In §3 we construct a sequence of equations whose unique solutions tend (uniformly in compact sets) to $\inf_u C^u(x) = \mathcal{Q}(x)$, where the infimum is over all admissible controls. The optimum control is admissible and minimizes (within U)

$$[\mathcal{Q}_x' f(x, u) + k(x, u)].$$

In §4 we establish criteria of the stochastic Liapunov function type, which may be used as a sufficient condition for the optimality of some given

² It is convenient to use this assumption in lieu of other more specific ones, such as bounded k or conditions on f , σ assuring a limited "rate of growth" of the solution paths, or conditions assuring the existence of a u implying the ergodicity properties sufficient to guarantee $E_x^u k(x_t, u_t) \rightarrow K < \infty$.

(possible unbounded) control and cost function. Section 5 gives a simple linear example which, although elementary, is also apparently new. The results of §4 can easily be extended to right continuous strong Markov processes with infinite escape time.

Blackwell [1], [2] considered the discounted cost problem for Markov chains. The results of this paper are based on a combination of the methods of Fleming [3], [4] and Kushner [5], [6], [7]. The continuous time discounted cost problem in an unbounded region does not appear to have been previously discussed. See also the remark following Theorem 1.

3. The solution to the discounted cost problem. The following lemma of Fleming [4] is required.

LEMMA 1. Let $F(x, u)$ be a real-valued function of the vectors x and u . Let $F_{u_i u_j}(x, u)$ be continuous and

$$|F_u(x + y, u) - F_u(x, u)| \leq M |y|, \quad M < \infty, \\ \sum_{i,j} F_{u_i u_j}(x, u) \mu_i \mu_j \geq \gamma |\mu|^2, \quad \gamma > 0 \quad \text{and any control vector } \mu,$$

where F_u is the gradient of F with respect to u . Let $\bar{u}(x)$ be the unique $u(x)$ with values in the compact convex set U which minimizes $F(x, u)$. Then $\bar{u}(x)$ is uniformly Lipschitz and

$$|\bar{u}(x + y) - \bar{u}(x)| \leq M\gamma^{-1} |y|.$$

See [4] for the proof.

LEMMA 2. Suppose the assumptions on f_i , a_{ij} , k , U and β of §1. Let u be admissible with corresponding process x_t . Define $T_n = \inf \{t: |x| \geq n\}$ and $S_n = \{x: |x| \leq n\}$, $n < \infty$. Then, if $x = x_0 \in S_n$, we have $E_x^u T_n \leq M_n < \infty$, and the unique solution W of (7) is (8):

$$\Lambda(W) + W_x' f + k = 0, \\ (7) \quad W(\partial S_n) = q(x), \quad \text{Hölder continuous,}$$

$$\Lambda(W) \equiv \sum_{i,j} a_{ij} W_{x_i x_j} - \beta W;$$

$$(8) \quad W(x) = C_n^u(x) \equiv E_x^u \int_0^{T_n} e^{-\beta t} k(x_t, u_t) dt + E_x^u e^{-\beta T_n} q(x_{T_n}).$$

$W(x)$ has Hölder continuous second derivatives and $T_n \rightarrow \infty$ w.p. 1 as $n \rightarrow \infty$.

Proof. a_{ij} , f_i , and k are uniformly bounded and satisfy a uniform Hölder condition in S_n . Alter these functions outside S_n so that they are uniformly bounded and satisfy the same uniform Hölder condition everywhere. (This does not effect the path behavior before T_n , w.p. 1.) Then $E_x^u T_n \leq M_n$ follows from Dynkin [9, Theorem 13.1], letting $V = 0$ and

$g = 1$. Equation (8) and the existence and uniqueness statement also follow from Dynkin [9, Theorems 13.16, 0.3]. $T_n \rightarrow \infty$ w.p. 1, since there is no finite escape time w.p. 1.

Lemma 3 is essentially a combination of two results of Fleming [3], [4] which will be needed.

LEMMA 3. *Suppose the conditions of §1 on f_i , a_{ij} , k , U and β . There is a unique solution to*

$$(9) \quad \begin{aligned} \Lambda(V) + \min_{u \in U} [V'_x f + k] &= 0, \\ V(\partial S_n) &= 0. \end{aligned}$$

$V(x)$ has Hölder continuous second derivatives.

The minimizing $u(x)$ is admissible and satisfies

$$V(x) = C_n^u(x) \equiv E_x^u \int_0^{T_n} e^{-\beta t} k(x_t, u_t) dt.$$

For any other admissible control w , $C_n^w(x) \geq C_n^u(x)$.

Proof. For a_{ij} , f_i and k Hölder continuous and bounded functions of their arguments in S_n , it follows from Fleming [3, Theorem 3] that (9) has a unique solution with Hölder continuous second derivatives. Then each V_{x_i} (component of V_x) has a bounded continuous derivative in S_n and, by the hypothesis on f and k , the conditions of Lemma 1 are satisfied with $F(x, u) = V'_x(x)f(x, u) + k(x, u)$. By Lemma 1, the minimizing function, $u(x)$, satisfies a Lipschitz condition in S_n . Hence, the x_t process corresponding to u is defined up to at least T_n , $E_x^u T_n < \infty$ (by Lemma 2), and V satisfies

$$\Lambda(V) + V'_x(x)f(x, u) + k(x, u) = 0.$$

Now, Lemma 2 gives the representation (10).

To prove the last statement, let w be admissible. Then $V'_x(x)f(x, w) + k(x, w)$ is bounded and Hölder continuous in S_n and, by the definition of u , we have

$$V'_x(x)f(x, u) + k(x, u) \leq V'_x(x)f(x, w) + k(x, w)$$

in S_n . Then,

$$\Lambda(V) + V'_x(x)f(x, w) \equiv -k_1(x, w) \geq -k(x, w)$$

and, by Lemma 2,

$$\begin{aligned} C_n^u(x) = V(x) &= E_x^u \int_0^{T_n(w)} e^{-\beta t} k_1(x_t, w_t) dt \\ &\leq E_x^u \int_0^{T_n(w)} e^{-\beta t} k(x_t, w_t) dt = C_n^w(x). \end{aligned}$$

Denote the solution to (9) by $V_n(x)$. $V_n(x)$ is defined in S_n only. We now construct a sequence $V_n(x)$, $n \rightarrow \infty$, and show that $\lim_n V_n(x) = \underline{C}(x)$, where $\underline{C}(x)$ satisfies

$$\Lambda(\underline{C}) + \min_{u \in U} [\underline{C}'_x(x)f(x, u) + k(x, u)] = 0.$$

The minimizing u is an admissible optimum control for (1), (3). Essentially $\underline{C}(x)$ is the smallest solution to $\Lambda(V) + [V'_x f + k] = 0$.

THEOREM 1 (Existence of an admissible optimal control). *Suppose the conditions of §1 on f_i , a_{ij} , U , β and k . The sequence $V_n(x) \uparrow V(x) = \underline{C}(x) < \infty$, where $V(x)$ has Hölder continuous second derivatives in any compact set. There is an admissible control, u , which minimizes (over $w(x) \in U$) $V'_x(x)f(x, w) + k(x, w)$ and is optimal for the problem (1), (3), and also*

$$\Lambda(V) + V'_x(x)f(x, u) + k(x, u) = 0.$$

Proof. Let $T_n(w) = \inf \{t: |x_t| \geq n\}$, where x_t corresponds to control w ; we use also $T_n = T_n(u)$. In S_n , $V_{n+1}(x) \geq V_n(x)$, since otherwise, using u_{n+1} (until $T_n(u_{n+1})$) yields, for some $x \in S_n$,

$$\begin{aligned} V_n(x) &\equiv C_n^{u_n}(x) = E_x^{u_n} \int_0^{T_n(u_n)} e^{-\beta t} k(x_t, u_{nt}) dt \\ &> E_x^{u_{n+1}} \int_0^{T_{n+1}(u_{n+1})} k(x_t, u_{n+1,t}) e^{-\beta t} dt \\ &\geq E_x^{u_{n+1}} \int_0^{T_n(u_{n+1})} e^{-\beta t} k(x_t, u_{n+1,t}) dt = C_n^{u_{n+1}}(x), \end{aligned}$$

which contradicts Lemma 3. Also $\underline{C}(x) < \infty$ by the hypothesis (see §1) that $C^w(x) < \infty$ for some admissible w . Since $C_n^{u_n}(x) \leq C^w(x)$, we have $C_n^{u_n}(x) = V_n(x) \uparrow V(x) \leq C^w(x) < \infty$.

Fix n , let $m \geq n + 1$. The interior estimates of Cordes [8, p. 303] for the solution of the elliptic equation $\Lambda(W) + W'_x f + k = 0$ apply to our problem in S_n and yield, for some α , $0 < \alpha < 1$, and $K < \infty$,

$$\begin{aligned} (10) \quad &\max_{x \in S_n} |V_m(x)| + \max_{x \in S_n} |V_x(x)| + \sup_{x, y \in S_n} \frac{|V_x(x+y) - V_x(y)|}{|y|^\alpha} \\ &\leq K \{ \max_{x \in S_{n+1}} |k(x, u_m(x))| + \max_{x \in S_{n+1}} |V_m(x)| \}. \end{aligned}$$

By (10), the families³ $\{V_m\}$ and $\{V_{mx}\}$ are equicontinuous. Thus $V_m \rightarrow V$ uniformly on S_n . There is a subsequence, indexed by m , so that $V_{mx} \rightarrow V_x$

³ V_{mx} is the gradient of V_m with respect to x . $V_{mx_i x_j}$ is the second partial derivative of V_m with respect to x_i, x_j .

uniformly on S_n (Ascoli's theorem). V_x is also Hölder continuous with Hölder exponent α and has partial derivatives almost everywhere (Lebesgue measure) in S_n . Furthermore, $V_{mx_{ij}} \rightarrow V_{x_{ij}}$ almost everywhere on S_n since (for the subsequence) $V_{mx} \rightarrow V_x$ uniformly there. At this time, define $V_{x_{ij}}$ arbitrarily on the null set.

Define $G(V_x, x, w) = V_x'(x)f(x, w) + k(x, w)$ and define $u(x) = u$ as the function w , with values in U , minimizing $G(V_x, x, w)$. u is Hölder continuous in S_n and, hence, so is $G_1(x) = G(V_x, x, u)$.

The following calculations are all in S_n . Let m index the abovementioned subsequence. Since u_m minimizes $G(V_{mx}, x, w)$,

$$\Lambda(V_m) + G(V_{mx}, x, u) \leq \Lambda(V_m) + G(V_{mx}, x, u_m) = 0$$

and

$$\Lambda(V) - \Lambda(V - V_m) + G(V_{mx}, x, u) \geq 0.$$

But $V_{mx_{ij}} \rightarrow V_{x_{ij}}$ almost everywhere and $V_{mx} \rightarrow V_x$ uniformly. Thus, almost everywhere

$$\Lambda(V) + G(V_x, x, u) \geq 0.$$

Also, by the definition of u , and use of $\Lambda(V_m) + G(V_{mx}, x, u_m) = 0$,

$$\Lambda(V) + G(V_x, x, u) \leq \Lambda(V - V_m) + [V_x - V_{mx}]'f(x, u_m) \rightarrow 0$$

almost everywhere. Thus,

$$(11) \quad \Lambda(V) + V_x'(x)f(x, u) + k(x, u) = 0$$

almost everywhere in S_n . Since the last two terms of (11) are Hölder continuous in S_n , a version of $V_{x_{ij}}$ can be chosen so that $V_{x_{ij}}$ is also Hölder continuous in S_n . Thus, V has Hölder continuous second derivatives in S_n and, by Lemma 1 (where $F(x, u) = V_x'(x)f(x, u) + k(x, u)$), u is uniformly Lipschitz in each S_n , hence it is admissible.

Note that $k \geq 0$ and $V_m(\partial S_m) = 0$ imply that $V_m(x) \geq 0$. Finally, the evaluation

$$\begin{aligned} V(x) &= E_x^u V(x_{T_n})e^{-\beta T_n} + E_x^u \int_0^{T_n} e^{-\beta t} k(x_t, u_t) dt \\ &= E_x^u \int_0^\infty e^{-\beta t} k(x_t, u_t) dt + \lim E_x^u V(x_{T_n})e^{-\beta T_n} \\ &= \lim_m V_m(x) \leq \mathcal{Q} \end{aligned}$$

implies that $V(x) = \mathcal{Q}$ and that u is optimal. (Note that the calculation implies that $E_x^u V(x_{T_n})e^{-\beta T_n} \rightarrow 0$.)

Remarks and extensions. The sets S_n could be any increasing collection

with smooth boundaries. Also, some smooth boundary conditions $V_n(\partial S_n) = q_n(x)$ could be imposed, provided $q_n(x)$ does not grow too fast with n . In particular, uniformly bounded q_n would yield the result of Theorem 1. It is not difficult to show that u is also optimal with respect to a class of nonanticipative controls (see Fleming [4] and Kushner [7] for the type of argument which is involved). The strong conditions (5), (6) are used only to insure that there is a well-defined process corresponding to each computed control. If one supposes that there is a suitable process and version of (8) corresponding to the differential generator (2) with less smooth coefficients (say, locally bounded and measurable), then a more general partial differential equation argument may be carried through. The problem with lack of smooth coefficients is strictly one of the probabilistic interpretation of the results.

Finally, the type of argument used in Theorem 1 may be applied to a general class of control problems, say, where the cost is

$$E_x^u \int_0^\tau k(x_t, u_t) dt, \quad k \geq 0,$$

where τ is the first time of arrival at a suitable set. The method, requiring the taking of a limit of a sequence of suitably truncated problems, is exactly that of Theorem 1.

4. Sufficient conditions for the optimality of a given control. If the ellipticity condition (4) is not satisfied, or if U is not bounded, the foregoing construction is not valid. Nevertheless, some sufficiency conditions for optimality can be given; if a function $V(x)$ and control $u(x)$ are given, we give a sufficient condition for the optimality of u , with respect to a given class of controls, and for the verification of $V(x) = C^u(x)$. Define $u = u(x)$ to be admissible if $u(x) \in U$, where U is an arbitrary constraint set, $u(x)$ is locally Lipschitz, and the escape time of the x_t process corresponding to u is infinite w.p. 1. (When U was bounded, the latter condition was guaranteed by the conditions on f_i and a_{ij} or σ_{ij} .)

LEMMA 4. *Let w be admissible in the sense of §4, and let f_i and σ_{ij} satisfy a local Lipschitz condition in their argument,⁴ and let k be continuous in its arguments. Let $V(x)$ have continuous second derivatives, and let τ be any nonanticipative finite-valued random time. If*

$$(12) \quad L^w V(x) \geq \beta V(x) - k(x, w), \quad k \geq 0, \quad \beta \geq 0,$$

then

⁴ Recall that, by hypothesis, the admissibility of w implies that the escape time is infinite.

$$(13) \quad V(x) \leq E_x^w V(x_\tau) e^{-\beta\tau} + E_x^w \int_0^\tau e^{-\beta s} k(x_s, w_s) ds.$$

If (12) is an equality, then so is (13).

Proof. Let $T_n = \inf \{t: |x_t| \geq n\}$. Alter the process for $t \geq T_n$ by defining f , σ and w outside of S_n so that they are uniformly Lipschitz and bounded. Alter V outside of S_n so that it is bounded and has bounded and continuous first and second derivatives. Now apply Itô's lemma (see [11, Theorem 2.5]) to $g(x, t) = e^{-\beta t} V(x)$ and the altered process and altered V . (The paths of the altered process are those of x_t w.p. 1, for $t < T_n$.) Then, for any finite-valued nonanticipative random time τ satisfying $\tau < T_n$, Itô's lemma and (12) yield w.p. 1 (in the most common version of Itô's lemma, τ (below) is a running time variable, not a random variable, but for τ nonanticipative and finite, the following equation is actually a special case of the general version of Itô's lemma given by Skorokhod [11])

$$\begin{aligned} V(x_\tau) e^{-\beta\tau} &= V(x) + \int_0^\tau e^{-\beta t} V_x'(x_t) dz_t \\ &\quad + \int_0^\tau e^{-\beta t} L^w V(x_t) dt - \beta \int_0^\tau e^{-\beta t} V(x_t) dt \end{aligned}$$

or for τ finite, nonanticipative, and $\tau < T_n$ (since, for $\tau < T_n$, the paths of the altered process are those of the unaltered process w.p. 1, we may use E_x^w as the expectation operator),

$$(14) \quad V(x) \leq E_x^w V(x_\tau) e^{-\beta\tau} + E_x^w \int_0^\tau e^{-\beta t} k(x_t, u_t) dt.$$

Equality in (12) gives equality in (14). Since n is arbitrary and $T_n \rightarrow \infty$ w.p. 1, the lemma is proved.

Remark. In general, if x_t is a right continuous strong Markov process with no finite escape time (for control w), then the weak infinitesimal operator replaces L^w and Dynkin's formula [9, p. 133] is used in lieu of Itô's lemma.

By a similar argument one proves the following lemma (proof omitted).

LEMMA 5. Suppose the conditions of Lemma 4 except that

$$L^w V(x) \leq \beta V(x)$$

replaces (12). Then

$$E_x^w V(x_\tau) e^{-\beta\tau} \leq V(x)$$

for any finite-valued nonanticipative random time τ .

Theorems 2 and 3 give conditions under which a solution, obtained via a

formal application of dynamic programming, actually yields an optimal control.

THEOREM 2 (Optimality theorem). *Let the conditions on f_i , σ_{ij} , β and k hold and suppose that w and u are admissible in the sense of §4. Let the non-negative function $V(x)$ have continuous second derivatives with*

$$(15a) \quad L^u V(x) = \beta V(x) - k(x, u),$$

$$(15b) \quad L^w V(x) \geq \beta V(x) - k(x, w),$$

and suppose that

$$(16) \quad \lim_n E_x^w V(x_{\tau_n}) e^{-\beta \tau_n} = A^w(x) = 0,$$

where τ_n is any sequence of finite-valued nonanticipative random times (corresponding to control w) tending to ∞ w.p. 1. Then,

$$C^u(x) = E_x^u \int_0^\infty e^{-\beta t} k(x_t, u_t) dt \leq C^w(x).$$

The control u is optimal with respect to all admissible w satisfying (15b) and (16).

Remark 1. Equation (15b) is implied by

$$\begin{aligned} (L^w - L^u)V(x) &= V'_x(x)[f(x, w) - f(x, u)] \\ &\leq k(x, u) - k(x, w). \end{aligned}$$

Remark 2. A sufficient condition for (16) is given in Theorem 3.

Remark 3. Suppose that dynamic programming were formally applied; then the functional equation

$$(17) \quad \sum a_{ij} V_{x_i x_j} - \beta V + \min_u [V'_x f + k] = 0$$

is obtained. Equation (17) may not have a solution, or may not have a unique solution, or the solution may not give an admissible control. Theorems 2 and 3 give a sufficient condition for a given solution of (17) to yield an optimal control.

Proof. Since the escape time for u and w is infinite w.p. 1 and $k \geq 0$, then both $C^u(x)$ and $C^w(x)$ are defined (finite or not). By Lemma 4, the nonnegativity of k and V , and (16), we have

$$\begin{aligned} (18) \quad V(x) &= \lim_n E_x^u V(x_{\rho_n}) e^{-\beta \rho_n} + E_x^u \int_0^\infty e^{-\beta t} k(x_t, u_t) dt \geq C^u(x), \\ V(x) &\leq \lim_n E_x^w V(x_{\tau_n}) e^{-\beta \tau_n} + E_x^w \int_0^\infty e^{-\beta t} k(x_t, u_t) dt \leq C^w(x). \end{aligned}$$

ρ_n is a sequence of finite-valued nonanticipative random times tending to

infinity w.p. 1. Equation (18) implies $C^w(x) \geq V(x) \geq C^u(x)$, and the theorem is proved.

The criteria for (16) given by Theorem 3 are expected to be rather easy to use. Let $P(\cdot)$ be the probability measure and ω the generic variable of the sample space.

THEOREM 3. *Suppose the conditions on f_i , σ_{ij} and V of Theorem 2. Let w be admissible. Let $F(\lambda)$ be a nonnegative, twice continuously differentiable function of the real variable λ . If $F(\lambda)/\lambda \rightarrow \infty$ monotonically as $\lambda \rightarrow \infty$ and*

$$L^w F(V(x)) \leq \beta F(V(x)),$$

then $A^w(x) = 0$.

Proof. Let $F(m)/m \equiv g_m \rightarrow \infty$ as $m \rightarrow \infty$. Lemma 5 is valid for $G(x) = F(V(x))$ replacing $V(x)$ and yields

$$(19) \quad G(x) \geq E_x^w G(x_{\tau_n}) e^{-\beta \tau_n}$$

for any sequence of finite-valued nonanticipative $\tau_n \rightarrow \infty$. Then, writing $V_n = V(x_{\tau_n})$ and using (19),

$$\int_{(\omega: V_n \geq m)} V_n e^{-\beta \tau_n} dP \leq \frac{1}{g(m)} \int_{(\omega: V_n \geq m)} G(x_{\tau_n}) e^{-\beta \tau_n} dP \leq \frac{G(x)}{g(m)}.$$

Since, in

$$\int V_n e^{-\beta \tau_n} dP = \int_{(\omega: V_n \geq m)} V_n e^{-\beta \tau_n} dP + \int_{(\omega: V_n < m)} V_n e^{-\beta \tau_n} dP,$$

the second term on the right tends to zero as $n \rightarrow \infty$, and the first term is bounded uniformly by $G(x)/g(m)$, for arbitrary m , the lemma is proved.

Remark. As in [6] (where different problems are considered) the function $F(\lambda) = \lambda \log(A + \lambda)$ for large A is expected to be useful in applications of the type of criterion given in Theorem 3.

5. A simple linear example. We very briefly discuss a simple linear-quadratic cost problem. The problem with filtering may be treated analogously to the finite time filtering and control example in [7, Chap. 4].

THEOREM 4. *Let the values of $u(x)$ be unconstrained and*

$$C^u(x) = E_x^u \int_0^\infty e^{-\beta t} (x_t' M x_t + u_t' Q u_t) dt,$$

$$dx = (Ax + Bu) dt + \sigma dz,$$

where M , Q , A , B , and σ are constant, M and Q are symmetric and Q is nonsingular. Let the real parts of the eigenvalues of $A - \beta I/2$ be negative. Then there is an optimal admissible control

$$(20) \quad u(x) = -Q^{-1}B'V_x(x)/2 = -Q^{-1}B'Px,$$

where

$$(21) \quad V(x) = \mathcal{Q}(x) = x'Px + r,$$

where

$$(22) \quad \beta r = \sum_{i,j} a_{ij}(P_{ij} + P_{ji})$$

and P is the unique (positive definite and symmetric) solution to

$$(23) \quad (A - \beta I/2)'P + P'(A - \beta I/2) - P'(BQ^{-1}B')P + M = 0.$$

$u(x)$ is optimal with respect to at least all admissible w for which

$$(24) \quad E_x {}^w x_t' P x_t e^{-\beta t} \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Proof. First, we formally use dynamic programming, then assume a quadratic solution, then we prove the optimality of the obtained control. Thus write, by dynamic programming ($V(x)$ is the candidate for $\mathcal{Q}(x)$),

$$(25) \quad 0 = \sum a_{ij} V_{x_i x_j}(x) - \beta V(x) \\ + \min_u [(Ax + Bu)' V_x(x) + x' Mx + u' Qu].$$

By supposing (P is assumed to be symmetric and positive definite) that

$$V(x) = x'Px + r,$$

the unconstrained minimizing $u(x)$ is easily computed to be

$$(26) \quad u(x) = -Q^{-1}B'V_x(x)/2 = -Q^{-1}B'Px.$$

Substituting (26) into (25) yields that (25) has a solution of the desired quadratic type only if (23) has a symmetric positive definite solution. r is given by (22).

If $(A - \beta I/2)$ is asymptotically stable, a result of Wonham [10, pp. 11–12] yields that (23) has a unique solution of the desired type.

Now, the $u(x)$ of (20) is admissible and, by construction, for any w ,

$$L^u V(x) = -k(x, u) + \beta V(x),$$

$$L^w V(x) \geq -k(x, w) + \beta V(x).$$

Therefore, u is optimal with respect to any w for which (16) holds. Since (24) is a special case of (16), the theorem is proved.

REFERENCES

- [1] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Statist., 33 (1962), pp. 719–726.
- [2] ———, *Discounted dynamic programming*, Ibid., 36 (1965), pp. 226–235.

- [3] W. H. FLEMING, *Some Markovian optimization problems*, J. Math. Mech., 12 (1963), pp. 131-140.
- [4] ———, *Duality and a priori estimates in Markovian optimization problems*, J. Math. Anal. Appl., to appear.
- [5] H. J. KUSHNER, *Stochastic stability and the design of feedback controls*, Proc. Polytechnic Institute of Brooklyn Symposium on Systems Theory, Polytechnic Press, New York, 1965, pp. 177-196.
- [6] ———, *Sufficient conditions for the optimality of a stochastic control*, this Journal, 3 (1966), pp. 499-508.
- [7] ———, *Stochastic Stability and Control*, Academic Press, New York, 1967.
- [8] H. O. CORDES, *Über die erste Randwertaufgabe bei quasilinearen Differentialgleichungen zweiter Ordnung in mehr als zwei Variablen*, Math. Ann., 131 (1956), pp. 278-312.
- [9] E. B. DYNKIN, *Markov Processes*, Springer-Verlag, Berlin, 1965.
- [10] W. M. WONHAM, *Optimum stationary control of a linear system with state-dependent noise*, Tech. Rep. 66-2, Center for Dynamical Systems, Brown University, Providence, Rhode Island, 1966.
- [11] A. K. SKOROKHOD, *Studies in the Theory of Random Processes*, Addison-Wesley, Reading, Massachusetts, 1965.
- [12] W. M. WONHAM, *Stochastic problems in optimal control*, Proc. 1963 IEEE International Convention Record, Part 2, 1963, pp. 114-124.
- [13] H. J. KUSHNER, *Some problems and some recent results in stochastic control*, Proc. 1965 IEEE International Convention, Part 6, 1965, pp. 108-116.
- [14] ———, *On the status of optimal control and stability for stochastic systems*, Proc. 1966 IEEE International Convention, Part 6, 1966, pp. 143-151.

ON AN EXISTENCE THEOREM FOR OPTIMAL CONTROL*

TOGO NISHIURA†

Introduction. In the first part of his paper [1], McShane proves an existence theorem for optimal control (see §3 below for the statement of the theorem). The proof in [1] used very elementary facts about Lebesgue integration and, for this reason, is very accessible to many readers. The present paper concerns another proof of the same theorem. The main tools used are the most elementary properties of Banach lattices and $L_p[a, b]$ spaces. The proof is effected by considering very simple bilinear functionals on appropriate spaces. Section 1 deals with definitions and notations. In §2, we discuss a general problem which includes McShane's theorem as a special case. We prove the existence theorem in §3.

1. Some definitions and notations. Throughout this paper R^n will be Euclidean n -space and I will be the closed interval $[a, b]$. Also, A will denote a compact subset of R^{n+1} and U a locally compact subset of R^m . Clearly, closed subsets of R^m are locally compact but not conversely. For example, any open set in R^m is also locally compact.

Suppose X is a topological space. Then $C(X)$ will denote the collection of all continuous functions on X . Let $C^*(X) = \{\phi \in C(X) \mid \phi \text{ is bounded}\}$ and $C_0(X) = \{\phi \in C(X) \mid \phi \text{ has compact support}\}$.

By a *measure* \mathfrak{M} we mean a linear functional $\mathfrak{M}: C_0(U) \rightarrow R$ such that:

- (A) $\phi \in C_0(U)$ and $\phi \geq 0$ imply $\mathfrak{M}(\phi) \geq 0$;
- (B) $\sup\{|\mathfrak{M}(\phi)| \mid \phi \in C_0(U), |\phi| \leq 1\} < \infty$.

It is clear that a measure \mathfrak{M} can be extended to a positive linear functional on the collection of all continuous functions ψ on U for which $\sup\{\mathfrak{M}(\phi) \mid \phi \in C_0(U), \phi \leq |\psi|\} < \infty$ in such a manner that the monotone convergence theorem holds. That is, $\psi_k \downarrow 0$ implies $\mathfrak{M}(\psi_k) \rightarrow 0$. Thus, it is obvious that $\mathfrak{M}(\psi)$ is defined for every $\psi \in C^*(U)$.

A one-parameter family of measures $\mathfrak{M}[\cdot; t]$, $t \in I$, is called a *measurable control* if, for each $\psi \in C_0(I \times U)$, $\mathfrak{M}[\psi(t, \cdot); t]$ is Lebesgue measurable on I . It is clear that the extended real-valued function $\mathfrak{M}[\psi(t, \cdot); t]$, where $\psi \in C(I \times U)$ and $\psi \geq 0$, is Lebesgue measurable. Thus, if ψ_1 and ψ_2 are two nonnegative continuous functions on $I \times U$, then $\mathfrak{M}[(\psi_1 - \psi_2)(t, \cdot); t]$ is Lebesgue measurable provided $\mathfrak{M}[\psi_i(t, \cdot); t]$ is finite almost everywhere on I , $i = 1, 2$, and equals $\mathfrak{M}[\psi_1(t, \cdot); t] - \mathfrak{M}[\psi_2(t, \cdot); t]$.

* Received by the editors April 11, 1967, and in revised form July 21, 1967.

† Department of Mathematics, Wayne State University, Detroit, Michigan 48202. This research was supported in part by the National Science Foundation under Grants GP 3834 and GP 7676.

A Banach lattice $B(A \times U)$, $\|\cdot\|$ is said to satisfy *condition C_0* if

(a) $C^*(A \times U) \subset B(A \times U) \subset C(A \times U)$, and

(b) $C_0(A \times U)$ is dense in $B(A \times U)$.

(See [2, p. 235] for the definition of a Banach lattice.) The properties of Banach lattices we will use most frequently are:

(i) $\phi \in B(A \times U)$ implies $|\phi| \in B(A \times U)$,

(ii) $\phi_1, \phi_2 \in B(A \times U)$ and $|\phi_1| \leq |\phi_2|$ imply $\|\phi_1\| \leq \|\phi_2\|$.

If $\phi \in C^*(A \times U)$, then $|\phi| \leq \|\phi\|_\infty 1$, where 1 is the constant function on $A \times U$ and $\|\phi\|_\infty$ is the usual $\sup\{|\phi(x, u)| \mid (x, u) \in A \times U\}$. Hence we have, by condition (a), $\|\phi\| \leq \|\phi\|_\infty \|1\|$ for every $\phi \in C^*(A \times U)$. By using an equivalent norm if necessary, we may assume $\|\phi\| \leq \|\phi\|_\infty$ for every $\phi \in C^*(A \times U)$. With the aid of condition (b), we can prove $\phi_j \in B(A \times U)$ and $\phi_j \downarrow 0$ imply $\|\phi_j\| \rightarrow 0$. That is, the monotone convergence theorem holds.

Finally, throughout this paper we shall assume p and q are such that $1 \leq p \leq 2 \leq q \leq \infty$ and $1/p + 1/q = 1$.

2. Measurable controls and bilinear functionals. In this section we discuss closure properties of sequences of measurable controls. The discussion leads to a general theorem which yields McShane's existence theorem as a corollary.

2.1. Construction of limit bilinear functionals. We assume the following hypotheses to hold in this subsection:

(i) $B, \|\cdot\|$ is a normed vector space and C is a dense subspace.

(ii) $\|\cdot\|_c$ is a norm on C such that $\|\phi\|_c \geq \|\phi\|$ for $\phi \in C$ and $C, \|\cdot\|_c$ is separable.

(iii) $T_k, k = 1, 2, \dots$, is a sequence of bilinear functionals such that $T_k(\phi, g)$ is defined whenever (a) $\phi \in B$ and $g \in L_q[a, b]$ or (b) $\phi \in C$ and $g \in L_p[a, b]$. Furthermore, in case (a), $|T_k(\phi, g)| \leq M \|\phi\| \|g\|_q$; and in case (b), $|T_k(\phi, g)| \leq M \|\phi\|_c \|g\|_p$. The constant M is independent of k .

PROPOSITION 1. *Let Φ be a countable dense subset of $C, \|\cdot\|_c$ and D be the space spanned by Φ . Then there is a subsequence $T_{k_j}, j = 1, 2, \dots$, of $T_k, k = 1, 2, \dots$, such that $T_{k_j}(\phi, g)$ converges for each $\phi \in D$ and $g \in L_p[a, b]$ to $T(\phi, g)$ so that T is a bilinear functional on $D \times L_p[a, b]$ with $|T(\phi, g)| \leq M \|\phi\|_c \|g\|_p$ and $|T(\phi, g)| \leq M \|\phi\| \|g\|_q$ for $(\phi, g) \in D \times L_q[a, b]$.*

Proof. For each $\phi \in \Phi$ and $k = 1, 2, \dots$, $T_k(\phi, \cdot)$ is a linear functional on $L_p[a, b]$, $1 \leq p \leq 2$, such that $|T_k(\phi, g)| \leq M \|\phi\|_c \|g\|_p$. Hence, $T_k(\phi, \cdot)$ has a weakly convergent subsequence. Since Φ is countable, by the Cantor diagonalization process, there is a subsequence $T_{k_j}, j = 1, 2, \dots$, such that $T_{k_j}(\phi, g)$ converges, say to $T(\phi, g)$, for every $(\phi, g) \in \Phi \times L_p[a, b]$. It is now a straightforward argument to define $T(\phi, g)$ for $(\phi, g) \in D \times L_p[a, b]$

so that $T_{k_j}(\phi, g) \rightarrow T(\phi, g)$. The remainder of the proposition is easily proved since $L_p[a, b] \supset L_q[a, b]$.

If D is a dense subspace of C , $\|\cdot\|_c$, then D is also a dense subspace of C , $\|\cdot\|$ since $\|\phi\|_c \geq \|\phi\|$. Consequently, D is a dense subspace of B , $\|\cdot\|$. Hence, if T is a bilinear functional defined on $D \times L_p[a, b]$ so that $|T(\phi, g)| \leq M \|\phi\|_c \|g\|_p$ and $|T(\phi, g)| \leq M \|\phi\| \|g\|_q$ for $(\phi, g) \in D \times L_q[a, b]$, then T has a unique bilinear extension to $B \times L_q[a, b]$ so that $|T(\phi, g)| \leq M \|\phi\| \|g\|_q$ and a unique bilinear extension to $C \times L_p[a, b]$ so that $|T(\phi, g)| \leq M \|\phi\|_c \|g\|_p$. Also, these extensions will agree on $C \times L_q[a, b]$ since $\|\phi\|_c \geq \|\phi\|$ for $\phi \in C$.

PROPOSITION 2. *Let D be a dense subspace of C , $\|\cdot\|_c$. Let T be a bilinear functional such that $|T(\phi, g)| \leq M \|\phi\| \|g\|_q$ for $(\phi, g) \in B \times L_q[a, b]$ and $|T(\phi, g)| \leq M \|\phi\|_c \|g\|_p$ for $(\phi, g) \in C \times L_p[a, b]$. Finally, let T_{k_j} , $j = 1, 2, \dots$, be a subsequence of T_k , $k = 1, 2, \dots$, such that $T_{k_j}(\phi, g) \rightarrow T(\phi, g)$ for $(\phi, g) \in D \times L_p[a, b]$. Then $T_{k_j}(\phi, g) \rightarrow T(\phi, g)$ for $\phi \in B$.*

Proof. Let $\phi \in B$, $\bar{\phi} \in D$ and $g \in L_q[a, b]$. Then

$$\begin{aligned} |T(\phi, g) - T_{k_j}(\phi, g)| &\leq |T(\phi - \bar{\phi}, g)| + |T(\bar{\phi}, g) - T_{k_j}(\bar{\phi}, g)| \\ &\quad + |T_{k_j}(\bar{\phi} - \phi, g)| \\ &\leq 2M \|\phi - \bar{\phi}\| \|g\|_q + |T(\bar{\phi}, g) - T_{k_j}(\bar{\phi}, g)|. \end{aligned}$$

Next, let $\phi \in C$, $\bar{\phi} \in D$ and $g \in L_p[a, b]$. Then

$$|T(\phi, g) - T_{k_j}(\phi, g)| \leq 2M \|\phi - \bar{\phi}\|_c \|g\|_p + |T(\bar{\phi}, g) - T_{k_j}(\bar{\phi}, g)|.$$

We conclude this subsection with the following theorem.

THEOREM 1. *There is a bilinear functional T such that:*

- (i) *some subsequence T_{k_j} , $j = 1, 2, \dots$, converges pointwise to T ;*
- (ii) $|T(\phi, g)| \leq M \|\phi\| \|g\|_q$ for $(\phi, g) \in B \times L_q[a, b]$;
- (iii) $|T(\phi, g)| \leq M \|\phi\|_c \|g\|_p$ for $(\phi, g) \in C \times L_p[a, b]$;
- (iv) $T(\phi, \cdot)$ *is absolutely continuous with respect to Lebesgue measure for every $\phi \in B$. That is, for each $\phi \in B$ and $\epsilon > 0$, there is a $\delta > 0$ such that $0 \leq g \leq 1$ and $\int_a^b g \, dt < \delta$ imply $|T(\phi, g)| < \epsilon$.*

Proof. We need only prove (iv). For $q < \infty$, (ii) implies (iv). For $q = \infty$, consider $\phi \in B$, $\bar{\phi} \in C$ and $g \in L_\infty[a, b]$. Then

$$\begin{aligned} |T(\phi, g)| &\leq |T(\phi - \bar{\phi}, g)| + |T(\bar{\phi}, g)| \\ &\leq M \|\phi - \bar{\phi}\| \|g\|_\infty + M \|\bar{\phi}\|_c \|g\|_1. \end{aligned}$$

Thus, (iv) follows.

2.2. Construction of a limit curve. The hypotheses for this subsection are:

- (i) $B(A \times U)$, $\|\cdot\|$ is a Banach lattice satisfying condition C_0 .

(ii) $f^i: A \times U \rightarrow R, i = 0, 1, \dots, n$, are continuous functions and α is an integer, $-1 \leq \alpha \leq n$, such that

(a) $0 \leq i \leq \alpha$ implies $f^i \in B(A \times U)$;

(b) $\alpha < i \leq n$ implies $f^i \geq 0$.

(iii) For each positive integer $k, x_k: I \rightarrow A$ is a continuous mapping and \mathfrak{M}_k is a measurable control such that

$$x_k^i(t) - x_k^i(a) = \int_a^t \mathfrak{M}_k[f^i(x_k(\tau), \cdot); \tau] d\tau, \quad t \in I, \quad i = 0, 1, \dots, n.$$

$$(iv) \left\{ \int_a^b (\mathfrak{M}_k[|\phi(x_k(t), \cdot)|; t])^p dt \right\}^{1/p} \leq M \|\phi\| \text{ for each } \phi \in B(A \times U)$$

and $k = 1, 2, \dots$.

(v) There is $H \in L_q[a, b]$ such that $\mathfrak{M}_k[1; t] \leq H(t)$ for almost every t in $I, k = 1, 2, \dots$.

Since A is compact, there is a subsequence of $x_k(a), k = 1, 2, \dots$, which converges. Also, for each $i > \alpha, x_k^i$ is a uniformly bounded sequence of nondecreasing functions on I . Hence, by Helly's theorem, there is a subsequence $x_{k_j}^i$ which converges on I to a nondecreasing function, say x_0^i .

Let us now turn to those $i \leq \alpha$. For each $\phi \in B(A \times U)$ and $g \in L_q[a, b]$, we have a bilinear functional $T_k(\phi, g) = \int_a^b g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt$. Also, for each $\phi \in C_0(A \times U)$ and $g \in L_p[a, b]$, we have a bilinear functional T_k given by the same integral. It is easy to see that $|T_k(\phi, g)| \leq M \|\phi\| \|g\|_q$ in the first case, and $|T_k(\phi, g)| \leq \|\phi\|_\infty \|H\|_q \|g\|_p$ in the second case. Hence, with the aid of Theorem 1, we have the next theorem.

THEOREM 2. *There are an increasing sequence of positive integers $k_j, j = 1, 2, \dots$, a bilinear functional T , and a mapping $x_0: I \rightarrow A$ such that:*

(i) $x_{k_j} \rightarrow x_0$ pointwise;

(ii) $T_{k_j} \rightarrow T$ pointwise;

(iii) x_0^i is nondecreasing for each $i > \alpha$;

(iv) x_0^i is absolutely continuous for each $i \leq \alpha$;

(v) $|T(\phi, g)| \leq M_0 \|\phi\| \|g\|_q$ for $(\phi, g) \in B(A \times U) \times L_q[a, b]$;

(vi) $|T(\phi, g)| \leq M_0 \|\phi\|_\infty \|g\|_p$ for $(\phi, g) \in C_0(A \times U) \times L_p[a, b]$;

(vii) $T(\phi, \cdot)$ is absolutely continuous with respect to Lebesgue measure for each $\phi \in B(A \times U)$.

M_0 above can be chosen to be $M + \|H\|_q$.

2.3. Limit curves and limit bilinear functionals. The following hypotheses are to hold for this subsection:

(i) $B(A \times U), \|\cdot\|$ is a Banach lattice satisfying condition C_0 .

(ii) $x_k: I \rightarrow A, k = 1, 2, \dots$, is a sequence of continuous mappings converging pointwise to $x_0: I \rightarrow A$ which need not be continuous.

(iii) $\mathfrak{M}_k, k = 1, 2, \dots$, is a sequence of measurable controls such that:

$$(a) \left\{ \int_a^b (\mathfrak{M}_k[\phi(x_k(t), \cdot); t])^2 dt \right\}^{1/p} \leq M \|\phi\| \text{ for } \phi \in B(A \times U),$$

and

(b) there is $H \in L_q[a, b]$ such that $\mathfrak{M}_k[1; t] \leq H(t)$ almost everywhere on I .

(iv) If $T_k(\phi, g) = \int_a^b g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt$, then $T_k(\phi, g) \rightarrow T(\phi, g)$

as $k \rightarrow \infty$ for $(\phi, g) \in B(A \times U) \times L_q[a, b]$ and for $(\phi, g) \in C_0(A \times U) \times L_p[a, b]$.

PROPOSITION 3. $\phi \in C_0(A \times U)$ implies $\mathfrak{M}_k[\phi(x_0(t), \cdot); t]$ is Lebesgue measurable. $\phi \in C(A \times U)$ and $\phi \geq 0$ imply $\mathfrak{M}_k[\phi(x_0(t), \cdot); t]$ is Lebesgue measurable.

Proof. For $\phi \in C_0(A \times U)$ we have

$$\mathfrak{M}_k[\phi(x_0(t), \cdot); t] = \lim_{j \rightarrow \infty} \mathfrak{M}_k[\phi(x_j(t), \cdot); t]$$

since $\phi(x_j(t), \cdot) \rightarrow \phi(x_0(t), \cdot)$ uniformly on U for each t . Suppose $\phi \in C(A \times U)$, $\phi_j \geq 0$, $\phi_j \in C_0(A \times U)$ and $\phi_j \uparrow \phi$. Then

$$\mathfrak{M}_k[\phi(x_0(t), \cdot); t] = \lim_{j \rightarrow \infty} \mathfrak{M}_k[\phi_j(x_0(t), \cdot); t].$$

DEFINITION 1. For $\phi \in C_0(A \times U)$, let

$$\bar{T}_k(\phi, g) = \int_a^b g(t) \mathfrak{M}_k[\phi(x_0(t), \cdot); t] dt.$$

PROPOSITION 4. $\phi \in C_0(A \times U)$ and $g \in L_p[a, b]$ imply $|T_k(\phi, g) - \bar{T}_k(\phi, g)| \rightarrow 0$ as $k \rightarrow \infty$.

Proof. Let $\epsilon > 0$ be given. Then there is an $\eta > 0$ such that $\mu(E) < \eta$ implies $\int_E |gH| dt < \epsilon$. Next, $\phi \in C_0(A \times U)$ implies there is a $\delta > 0$ such that $|x' - x''| < \delta$ implies $\|\phi(x', \cdot) - \phi(x'', \cdot)\|_\infty < \epsilon$. Since $x_k \rightarrow x_0$ pointwise, by Egorov's theorem, there are a set E such that $\mu(E) < \eta$ and an integer K such that $|x_k(t) - x_0(t)| < \delta$ uniformly on $I \setminus E$ for all $k > K$. Then for $k > K$, we have

$$\begin{aligned} & |T_k(\phi, g) - \bar{T}_k(\phi, g)| \\ & \leq \int_a^b |g(t)| \mathfrak{M}_k[|\phi(x_k(t), \cdot) - \phi(x_0(t), \cdot)|; t] dt \\ & \leq \int_E |g(t)| 2 \|\phi\|_\infty H(t) dt + \int_{I \setminus E} |g(t)| \epsilon H(t) dt \\ & \leq (2\|\phi\|_\infty + \|gH\|_1) \epsilon. \end{aligned}$$

Proposition 4 shows the limit $\bar{T}(\phi, g) = \lim_{k \rightarrow \infty} \bar{T}_k(\phi, g)$ exists for each $(\phi, g) \in C_0(A \times U) \times L_p[a, b]$ and $\bar{T}(\phi, g) = T(\phi, g)$. Hence, we may extend $\bar{T}(\phi, g)$ to each $(\phi, g) \in B(A \times U) \times L_q[a, b]$ and $\bar{T}(\phi, g) = T(\phi, g)$. Furthermore, if $\phi, \psi \in B(A \times U)$ and $\phi(x_0(t), u) = \psi(x_0(t), u)$ for all $(t, u) \in I \times U$, then $\bar{T}(\phi, g) = \bar{T}(\psi, g)$.

PROPOSITION 5. $\phi \geq 0$ and $g \geq 0$ imply $\bar{T}(\phi, g) \geq 0$.

Proof. This is obvious since $T_k(\phi, g) \geq 0$ for $\phi \geq 0$ and $g \geq 0$.

PROPOSITION 6. Suppose $\phi \in C(A \times U)$ and $\phi \geq 0$. Let $\phi_j = \phi \wedge j$ and $g \in L_\infty[a, b]$, $g \geq 0$. Then

$$\begin{aligned} \lim_{j \rightarrow \infty} \bar{T}(\phi_j, g) &\leq \lim_{k \rightarrow \infty} \int_a^b g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt \\ &\leq \|g\|_\infty \lim_{k \rightarrow \infty} \int_a^b \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt. \end{aligned}$$

Proof. $g(t) \mathfrak{M}_k[\phi_j(x_k(t), \cdot); t] \leq g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t]$ for every $t \in I$ and every j . Also, $\bar{T}(\phi_j, g) \leq \bar{T}(\phi_{j+1}, g)$. Hence,

$$\begin{aligned} \lim_{j \rightarrow \infty} \bar{T}(\phi_j, g) &= \lim_{j \rightarrow \infty} T(\phi_j, g) = \lim_{j \rightarrow \infty} \lim_{k \rightarrow \infty} T_k(\phi_j, g) \\ &\leq \lim_{j \rightarrow \infty} \lim_{k \rightarrow \infty} \int_a^b g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt \\ &= \lim_{k \rightarrow \infty} \int_a^b g(t) \mathfrak{M}_k[\phi(x_k(t), \cdot); t] dt. \end{aligned}$$

PROPOSITION 7. Let \bar{t} be a point of continuity of x_0 , $\epsilon > 0$ and $\phi \in C_0(A \times U)$. Then there exists $\bar{h} > 0$ such that $g \in L_p[a, b]$ and $\{t \mid g(t) \neq 0\} \subset [\bar{t} - \bar{h}, \bar{t} + \bar{h}]$ imply $|\bar{T}(\phi, g)| \leq (\|\phi(x_0(\bar{t}), \cdot)\|_\infty + \epsilon) \|gH\|_1$.

Proof. Since $\phi \in C_0(A \times U)$, ϕ is uniformly continuous. Hence, there is $\delta > 0$ such that $|x' - x''| < \delta$ implies $|\phi(x', u) - \phi(x'', u)| < \epsilon$ for all $u \in U$. Since \bar{t} is a point of continuity of x_0 , there is an $\bar{h} > 0$ such that $|t - \bar{t}| < \bar{h}$ implies $|x_0(t) - x_0(\bar{t})| < \delta$. Let $g \in L_p[a, b]$ and $\{t \mid g(t) \neq 0\} \subset [\bar{t} - \bar{h}, \bar{t} + \bar{h}]$. Then

$$\begin{aligned} |\bar{T}(\phi, g)| &= \lim_{k \rightarrow \infty} |\bar{T}_k(\phi, g)| \\ &\leq \lim_{k \rightarrow \infty} \int_a^b |g(t)| \mathfrak{M}_k[\phi(x_0(t), \cdot); t] dt \\ &\leq \lim_{k \rightarrow \infty} (\|\phi(x_0(\bar{t}), \cdot)\|_\infty + \epsilon) \int_a^b |g(t)| \mathfrak{M}_k[1; t] dt \\ &\leq (\|\phi(x_0(\bar{t}), \cdot)\|_\infty + \epsilon) \|gH\|_1. \end{aligned}$$

2.4. Construction of a measurable control. The hypotheses of this subsection are:

- (i) $B(A \times U)$, $\|\cdot\|$ is a Banach lattice satisfying condition C_0 .
- (ii) $x_0 : I \rightarrow A$ is not necessarily continuous.
- (iii) $H \in L_q[a, b]$, $H \geq 0$ and H real-valued.
- (iv) \bar{T} is a bilinear functional such that:

$$\begin{aligned} |\bar{T}(\phi, g)| &\leq M \|\phi\| \|g\|_q \quad \text{for } (\phi, g) \in B(A \times U) \times L_q[a, b], \\ |\bar{T}(\phi, g)| &\leq M \|\phi\|_\infty \|g\|_p \quad \text{for } (\phi, g) \in C_0(A \times U) \times L_p[a, b], \\ \bar{T}(\phi, g) &\geq 0 \quad \text{for } \phi \geq 0, g \geq 0. \end{aligned}$$

(v) There is a set E of measure zero such that $\bar{t} \notin E$, $\epsilon > 0$ and $\phi, \psi \in C_0(A \times U)$ imply there is an $\bar{h} > 0$ so that $|\bar{T}(\phi - \psi, g)| \leq (\|\phi(x_0(\bar{t}), \cdot) - \psi(x_0(\bar{t}), \cdot)\|_\infty + \epsilon) \|gH\|_1$ whenever $g \in L_p[a, b]$ and $\{t \mid g(t) \neq 0\} \subset [\bar{t} - \bar{h}, \bar{t} + \bar{h}]$.

DEFINITION 2. Let $t \in I$ and $h > 0$ and define $g_{t,h}(\tau)$ to be h^{-1} for $t \leq \tau \leq t + h$ and 0 for all other τ . Let $\mathfrak{M}[\phi, t] = \lim_{h \rightarrow 0+} \bar{T}(\phi, g_{t,h})$, where $\phi \in C_0(A \times U)$. Since $\bar{T}(\phi, \cdot)$ is absolutely continuous with respect to Lebesgue measure (see proof of Theorem 1), the limit exists for almost all $t \in I$. Finally, let $E' = \{t \mid \lim_{h \rightarrow 0+} \|g_{t,h} H\|_1 > H(t)\}$. Clearly, $\mu(E') = 0$.

PROPOSITION 8. Let $t \notin E \cup E'$ and $\phi, \psi \in C_0(A \times U)$ for which $\phi(x_0(t), \cdot) = \psi(x_0(t), \cdot)$. Then we have $\mathfrak{M}[\phi, t]$ exists if and only if $\mathfrak{M}[\psi, t]$ exists, and the limits are equal when they exist. Furthermore, $|\mathfrak{M}[\phi, t]| \leq \|\phi(x_0(t), \cdot)\|_\infty H(t)$.

Proof. By hypothesis (v) of this subsection, for each $\epsilon > 0$ there is an $\bar{h} > 0$ such that $|\bar{T}(\phi, g_{t,h}) - \bar{T}(\psi, g_{t,h})| \leq \epsilon \|g_{t,h} H\|_1$ for $0 < h < \bar{h}$. Hence the first part of the proposition follows. The second part follows from (v) again by choosing $\psi = 0$.

DEFINITION 3. Each $\eta \in C_0(U)$ can be considered a member of $C_0(A \times U)$. Let $\mathfrak{M}_0[\eta; t] = \mathfrak{M}[\eta; t]$.

PROPOSITION 9. There is a set Z of measure zero such that $t \notin Z$ implies $\mathfrak{M}[\phi; t]$ exists for all $\phi \in C_0(A \times U)$ and $\mathfrak{M}[\phi; t] = \mathfrak{M}_0[\phi(x_0(t), \cdot); t]$.

Proof. Let D be a countable dense subset of $C_0(U)$, $\|\cdot\|_\infty$. There is a set Z' of measure zero such that $t \notin Z'$ implies $\mathfrak{M}_0[\eta; t]$ exists for all $\eta \in D$. Let $Z = E \cup E' \cup Z'$. Now consider $t \notin Z$, $\phi \in C_0(A \times U)$ and let $\{\eta_j\}$ be a sequence in D such that $\|\phi(x_0(t), \cdot) - \eta_j\|_\infty \rightarrow 0$. Clearly, $\{\eta_j\}$ is a Cauchy sequence in $C_0(U)$. By Proposition 8, we have

$$|\mathfrak{M}_0[\eta_j; t] - \mathfrak{M}_0[\eta_k; t]| \leq \|\eta_j - \eta_k\|_\infty H(t).$$

By hypothesis (v), we have

$$\begin{aligned} \mathfrak{M}_0[\eta_j; t] - \|\eta_j - \phi(x_0(t), \cdot)\|_\infty H(t) &\leq \lim_{h \rightarrow 0+} \bar{T}(\phi, g_{t,h}) \leq \overline{\lim}_{h \rightarrow 0+} \bar{T}(\phi, g_{t,h}) \\ &\leq \mathfrak{M}_0[\eta_j; t] + \|\eta_j - \phi(x_0(t), \cdot)\|_\infty H(t). \end{aligned}$$

Hence, $\mathfrak{M}[\phi; t]$ exists. The existence of $\mathfrak{M}_0[\phi(x_0(t), \cdot); t]$ and the equality $\mathfrak{M}[\phi; t] = \mathfrak{M}_0[\phi(x_0(t), \cdot); t]$ follow from Proposition 8.

PROPOSITION 10. *Let $t \notin Z$. Then,*

(a) $\mathfrak{M}_0[\cdot; t]$ is a positive linear functional on $C_0(U)$, and

(b) $\eta \in C_0(U)$ and $|\eta| \leq 1$ imply $|\mathfrak{M}_0[\eta; t]| \leq H(t)$.

Proof. Part (a) is obvious since $\bar{T}(\eta, g_{t,h}) \geq 0$ if $\eta \geq 0$ and differentiation is a linear operator. By Propositions 8 and 9, we have (b).

For $t \in Z$, let $\mathfrak{M}_0[\cdot; t]$ be an arbitrary measure. The next proposition will complete the proof that \mathfrak{M}_0 is a measurable control.

PROPOSITION 11. *Suppose $\psi \in C_0(I \times U)$. Then $\mathfrak{M}_0[\psi(t, \cdot); t]$ is Lebesgue measurable.*

Proof. Suppose $\psi \in C_0(I \times U)$ and let $\epsilon > 0$. There exists a $\delta > 0$ such that $|t - t'| < \delta$ implies $|\psi(t, u) - \psi(t', u)| < \epsilon$ for all $u \in U$. Let $a = t_0 < t_1 < \dots < t_k = b$ such that $|t_i - t_{i-1}| < \delta$, $i = 1, \dots, k$. Then we have for $t_{i-1} \leq t \leq t_i$ and $u \in U$, $\psi(t_i, u) - \epsilon \leq \psi(t, u) \leq \psi(t_i, u) + \epsilon$. Let χ_i be the characteristic function of the interval $[t_{i-1}, t_i]$, $i = 1, 2, \dots, k$. Then for each i , $\chi_i(t)\mathfrak{M}_0[\psi(t_i, \cdot); t]$ is measurable and

$$\begin{aligned} \chi_i(t)\{\mathfrak{M}_0[\psi(t_i, \cdot); t] - \epsilon H(t)\} &\leq \chi_i(t)\mathfrak{M}_0[\psi(t, \cdot); t] \\ &\leq \chi_i(t)\{\mathfrak{M}_0[\psi(t_i, \cdot); t] + \epsilon H(t)\} \end{aligned}$$

almost everywhere on I . Hence, for each $\epsilon > 0$, there is a measurable function F_ϵ such that $F_\epsilon(t) - \epsilon H(t) \leq \mathfrak{M}_0[\psi(t, \cdot); t] \leq F_\epsilon(t) + \epsilon H(t)$ almost everywhere on I . The proposition follows.

We conclude this subsection with the following theorem.

THEOREM 3. *There exists a measurable control \mathfrak{M}_0 such that:*

- (i) $\mathfrak{M}_0[1; t] \leq H(t)$ almost everywhere on I ,
- (ii) $\bar{T}(\phi, g) = \int_a^b g(t) \mathfrak{M}_0[\phi(x_0(t), \cdot); t] dt$ for $\phi \in B(A \times U)$, and
- (iii) if $\phi \in C(A \times U)$, $\phi \geq 0$, $\phi_j = \phi \wedge j$, $g \in L_\infty[a, b]$ and $g \geq 0$, then

$$\lim_{j \rightarrow \infty} \bar{T}(\phi_j, g) = \int_a^b g(t) \mathfrak{M}_0[\phi(x_0(t), \cdot); t] dt.$$

Proof. Part (i) follows easily from Proposition 10(b).

We will next prove part (ii). If $\phi \in C_0(A \times U)$, then clearly $\bar{T}(\phi, g) = \int_a^b g(t) \mathfrak{M}_0[\phi(x_0(t), \cdot); t] dt$ for $g \in L_p[a, b]$. Suppose $\psi \in C(A \times U)$, $\psi \geq 0$ and $\phi_j \in C_0(A \times U)$, $0 \leq \phi_j \leq \psi$ and $\phi_j \uparrow \psi$. Then the extended real-valued function $\mathfrak{M}_0[\psi(x_0(t), \cdot); t]$ is equal to $\lim_{j \rightarrow \infty} \mathfrak{M}_0[\phi_j(x_0(t), \cdot); t]$. Also, $\mathfrak{M}_0[\phi_j(x_0(t), \cdot); t] \leq \mathfrak{M}_0[\phi_{j+1}(x_0(t), \cdot); t]$, $t \in I$. Hence, for $g \in L_q[a, b]$, $g \geq 0$, $\psi \in B(A \times U)$, $\psi \geq 0$, we have

$$\bar{T}(\psi, g) = \lim_{j \rightarrow \infty} \bar{T}(\phi_j, g)$$

$$\begin{aligned}
&= \lim_{j \rightarrow \infty} \int_a^b g(t) \mathfrak{M}_0[\phi_j(x_0(t), \cdot); t] dt \\
&= \int_a^b g(t) \mathfrak{M}_0[\psi(x_0(t), \cdot); t] dt.
\end{aligned}$$

Hence, $\mathfrak{M}_0[\psi(x_0(t), \cdot); t]$ is finite almost everywhere. Thus, we infer $\mathfrak{M}_0[\psi(x_0(t), \cdot); t]$ is defined almost everywhere for $\psi \in B(A \times U)$ and

$$\mathfrak{M}_0[\psi(x_0(t), \cdot); t] = \mathfrak{M}_0[\psi^+(x_0(t), \cdot); t] - \mathfrak{M}_0[\psi^-(x_0(t), \cdot); t]$$

almost everywhere on I . Clearly, the condition $\psi \geq 0$ and $g \geq 0$ can now be eliminated and (ii) holds.

The proof of part (iii) is a variant of that of (ii).

2.5. Statement of the main theorem.

THEOREM 4. *Suppose the following hypotheses hold:*

- (i) $B(A \times U)$, $\|\cdot\|$ is a Banach lattice satisfying condition C_0 .
- (ii) $f^i: A \times U \rightarrow R$, $i = 0, 1, \dots, n$, are continuous functions and α is an integer, $-1 \leq \alpha \leq n$, such that:

(a) $0 \leq i \leq \alpha$ implies $f^i \in B(A \times U)$,

(b) $\alpha < i \leq n$ implies $f^i \geq 0$.

- (iii) For each positive integer k , $x_k: I \rightarrow A$ is a continuous mapping and \mathfrak{M}_k is a measurable control such that

$$x_k^i(t) - x_k^i(a) = \int_a^t \mathfrak{M}_k[f^i(x_k(\tau), \cdot); \tau] d\tau,$$

$$t \in I, \quad i = 0, \dots, n$$

$$(iv) \left\{ \int_a^b (\mathfrak{M}_k[\|\phi(x_k(t), \cdot)\|; t])^p dt \right\}^{1/p} \leq M \|\phi\| \text{ for each } \phi \in B(A \times U)$$

and $k = 1, 2, \dots$.

- (v) There is an $H \in L_q[a, b]$ such that $\mathfrak{M}_k[1; t] \leq H(t)$ for almost every $t \in I$, $k = 1, 2, \dots$.

Then there are a mapping $x_0: I \rightarrow A$ (not necessarily continuous), a measurable control \mathfrak{M}_0 , and an increasing sequence of integers k_j , $j = 1, 2, \dots$, such that:

- (vi) $x_{k_j} \rightarrow x_0$ pointwise,
- (vii) $\mathfrak{M}_0[1; t] \leq H(t)$ almost everywhere on I ,
- (viii) $0 \leq i \leq \alpha$ implies

$$x_0^i(t) - x_0^i(a) = \int_a^t \mathfrak{M}_0[f^i(x_0(\tau), \cdot); \tau] d\tau,$$

and $\alpha < i \leq n$ implies

$$x_0^i(t) - x_0^i(a) \geq \int_a^t \mathfrak{M}_0[f^i(x_0(\tau), \cdot); \tau] d\tau.$$

Proof. Parts (vi) and (vii) are immediate from the above discussion. We prove (viii). Let χ_t be the characteristic function of the interval $[a, t]$. For $0 \leq i \leq \alpha$, we have

$$\begin{aligned} x_0^i(t) - x_0^i(a) &= T(f^i, \chi_t) = \bar{T}(f^i, \chi_t) \\ &= \int_a^b \chi_t(\tau) \mathfrak{M}_0[f^i(x_0(\tau), \cdot); \tau] d\tau \\ &= \int_a^t \mathfrak{M}_0[f^i(x_0(\tau), \cdot); \tau] d\tau. \end{aligned}$$

Let $\alpha < i \leq n$. Suppose $0 \leq \phi_j$ and $\phi_j \uparrow f^i$, where $\phi_j \in C_0(A \times U)$. Then by (vi) above and Theorem 3, we have

$$\begin{aligned} x_0^i(t) - x_0^i(a) &= \lim_{k \rightarrow \infty} (x_k^i(t) - x_k^i(a)) \\ &= \lim_{k \rightarrow \infty} \int_a^t \mathfrak{M}_k[f^i(x_k(\tau), \cdot); \tau] d\tau \\ &\geq \lim_{j \rightarrow \infty} \lim_{k \rightarrow \infty} \int_a^t \mathfrak{M}_k[\phi_j(x_k(\tau), \cdot); \tau] d\tau \\ &= \lim_{j \rightarrow \infty} \bar{T}(\phi_j, \chi_t) = \int_a^t \mathfrak{M}_0[f^i(x_0(\tau), \cdot); \tau] d\tau. \end{aligned}$$

The theorem is proved.

3. Proof of the existence theorem. We recall the following notations and definitions from [1] for the reader's convenience:

B is any subset of R^n ; U is a closed subset of R^m ; $J = [t_0, t_1]$; E is a closed subset of $R \times B \times R \times B$ and (t_0, x_0, t_1, x_1) denotes a point of E .

$e: E \rightarrow R$ is a continuous function. For each $i = 1, 2, \dots, n$, $f^i: R \times B \times U \rightarrow R$ is a continuous function.

A pair (\mathfrak{M}, J) is called a *relaxed control* if $\mathfrak{M}[\cdot; t]$ is a one parameter family of measures, $t \in R$, such that \mathfrak{M} is a measurable control on every interval I containing J and, for all t , $\mathfrak{M}[1; t] = \chi_J(t)$, where χ_J is the characteristic function of the interval J .

A triple $C = (\mathfrak{M}, x(\cdot), J)$ is called a *generalized curve* if $x: J \rightarrow B$ is a continuous mapping, (\mathfrak{M}, J) is a relaxed control and $x^i(t) = x^i(t_0) + \int_{t_0}^t \mathfrak{M}[f^i(\tau, x(\tau), \cdot); \tau] d\tau$, $t \in J$, $i = 1, 2, \dots, n$. A generalized curve C is called *admissible* if $\eta(C) = (t_0, x(t_0), t_1, x(t_1)) \in E$. \mathfrak{F} denotes the family of all admissible curves and $e_{\min} = \inf \{e(\eta(C)) \mid C \in \mathfrak{F}\}$.

A subset V of R^r is said to be *independent of x^j* if $(\bar{x}^1, \dots, \bar{x}^{j-1}, \bar{x}^j, \bar{x}^{j+1}, \dots, \bar{x}^r) \in V$ implies $(\bar{x}^1, \dots, \bar{x}^{j-1}, \bar{x}^j, \bar{x}^{j+1}, \dots, \bar{x}^r) \in V$ for all $\bar{x}^j \in R$.

Suppose A is a compact subset of $R \times B \subset R^{n+1}$ and $G \in C(A \times U)$ and $G \geq 0$. Let $\psi \in C(A \times U)$. We say ψ is of *slower growth than G uniformly on A* if, for each $\epsilon > 0$, there is a bounded subset U_ϵ of U such that $|\psi(t, x, u)| \leq \epsilon G(t, x, u)$ whenever $(t, x) \in A$ and $u \in U \setminus U_\epsilon$.

We now state the theorem.

THEOREM (McShane). *For each positive integer k , $C_k = (\mathfrak{M}_k, x_k(\cdot), J_k)$ is an admissible generalized curve. We suppose that the following hold:*

1. $\lim_{k \rightarrow \infty} e(\eta(C_k)) = e_{\min}$.
2. $\{(t, x_k(t)) \mid t \in J_k\} \subset A$, $k = 1, 2, \dots$, where A is a compact subset of $R \times B$.

3. *There is $G \in C(A \times U)$, $G \geq 0$, such that:*

- (a) $\int_{t_{0,k}}^{t_{1,k}} \mathfrak{M}_k[G(t, x_k(t), \cdot); t] dt \leq M < \infty$, $k = 1, 2, \dots$;
- (b) *1 is of slower growth than G uniformly on A ;*
- (c) *there is an integer α , $0 \leq \alpha \leq n$, such that:*
 - (i) *$1 \leq i \leq \alpha$ implies f^i is of slower growth than G uniformly on A , and $\alpha < i \leq n$ implies $f^i \geq 0$,*
 - (ii) *$f^j, j = 1, 2, \dots, n$, are independent of $x^i, \alpha < i \leq n$,*
 - (iii) *B is independent of x^i and E is independent of $x_1^i, \alpha < i \leq n$,*
 - (iv) *e is a nondecreasing function of $x_1^i, \alpha < i \leq n$.*

Then the family \mathcal{F} of admissible generalized curves contains a member $C_0 = (\mathfrak{M}_0, x_0(\cdot), J_0)$ such that $e(\eta(C_0)) = e_{\min}$.

Proof. Let $G_1 = G + 1$. Then 1 and $f^i, 1 \leq i \leq \alpha$, are of slower growth than G_1 uniformly on A . Let $\mathcal{G}(A \times U) = \{\phi \in C(A \times U) \mid |\phi| \leq \lambda G_1 \text{ for some } \lambda \geq 0\}$, and $\|\phi\| = \inf \{\lambda \mid \lambda \geq 0, |\phi| \leq \lambda G_1\}$. Then it is easy to see that $\mathcal{G}(A \times U), \|\cdot\|$ is a Banach lattice such that $C^*(A \times U) \subset \mathcal{G}(A \times U)$. Let $B(A \times U)$ be the norm closure of $C^*(A \times U)$. Then $B(A \times U), \|\cdot\|$ is a Banach lattice such that $C^*(A \times U) \subset B(A \times U)$. A simple calculation shows that 1 being of slower growth than G_1 uniformly on A yields $C_0(A \times U)$ dense in $B(A \times U), \|\cdot\|$. Hence, $B(A \times U), \|\cdot\|$ is a Banach lattice satisfying condition C_0 . Also, (i) of 3(c) above implies $f^i \in B(A \times U)$ for $1 \leq j \leq \alpha$.

Since A is compact, there is an interval $I = [a, b]$ such that $I \supset J_k$, $k = 1, 2, \dots$, and hence \mathfrak{M}_k is a measurable control on I such that $\mathfrak{M}_k[1; t] = \chi_{J_k}(t) \leq 1, t \in I$.

Next let $f^0(t, x, u) = 1$ on $R \times B \times U$ and $X_k = (X_k^0, X_k^1, \dots, X_k^n): I \rightarrow A$ be given by

$$X_k^i(t) = X_k^i(a) + \int_a^t \mathfrak{M}_k[f^i(X_k(\tau), \cdot); \tau] d\tau, \quad t \in I,$$

where $X_k^0(a) = t_{0,k}$ and $X_k^i(a) = x_k^i(t_{0,k}), i = 1, 2, \dots, n$. Clearly, X_k^i restricted to J_k is x_k^i for $i = 1, 2, \dots, n$ and $X_k^0(t) = t$ for $t \in J_k$.

Finally,

$$\begin{aligned} \int_a^b \mathfrak{M}_k[|\phi(X_k(t), \cdot)|; t] dt &\leq \int_a^b \mathfrak{M}_k[\|\phi\|(G(X_k(t), \cdot) + 1); t] dt \\ &\leq [M + b - a]\|\phi\| \end{aligned}$$

for $\phi \in B(A \times U)$. Hence, by choosing $p = 1$, we have the hypothesis of the main theorem verified.

Consequently, we have an increasing sequence of integers $\{k_j\}$, a measurable control \mathfrak{M}_0 and a mapping $X_0: I \rightarrow A$ such that:

- (i) $X_{k_j} \rightarrow X_0$ pointwise,
- (ii) $\mathfrak{M}_0[1; t] \leq 1$ on I ,
- (iii) $0 \leq i \leq \alpha$ implies

$$X_0^i(t) - X_0^i(a) = \int_a^t \mathfrak{M}_0[f^i(X_0(\tau), \cdot); \tau] d\tau,$$

and $\alpha < i \leq n$ implies

$$X_0^i(t) - X_0^i(a) \geq \int_a^t \mathfrak{M}_0[f^i(X_0(\tau), \cdot); \tau] d\tau.$$

Let $t_{0,0} = \lim_{j \rightarrow \infty} X_{k_j}^0(a) = \lim_{j \rightarrow \infty} t_{0,k_j}$ and $t_{1,0} = \lim_{j \rightarrow \infty} X_{k_j}^0(b) = \lim_{j \rightarrow \infty} t_{1,k_j}$ and let $J_0 = [t_{0,0}, t_{1,0}]$. It is a simple matter to see that $\mathfrak{M}_0[1; t] = \chi_{J_0}(t)$ for $t \in I$ since

$$X_0^0(t) = \begin{cases} t_{0,0} & \text{for } a \leq t \leq t_{0,0}, \\ t & \text{for } t_{0,0} \leq t \leq t_{1,0}, \\ t_{1,0} & \text{for } t_{1,0} \leq t \leq b. \end{cases}$$

Let

$$x_0^i(t) = x_0^i(t_{0,0}) + \int_{t_{0,0}}^t \mathfrak{M}_0[f^i(X_0(\tau), \cdot); \tau] d\tau, \quad t \in J_0,$$

where $x_0^i(t_{0,0}) = X_0^i(a)$, $i = 1, 2, \dots, n$. From (ii) and (iii) of hypothesis 3(c) we have

$$\begin{aligned} x_0^i(t) &= x_0^i(t_{0,0}) + \int_{t_{0,0}}^t \mathfrak{M}_0[f^i(\tau, x_0(\tau), \cdot); \tau] d\tau, \\ &\quad i = 1, 2, \dots, n, \end{aligned}$$

and $x_0: J_0 \rightarrow B$ is a continuous mapping. Since E is independent of x_1^i , $\alpha < i \leq n$, we have $C_0 = (\mathfrak{M}_0, x_0(\cdot), J_0)$ is an admissible generalized curve.

Finally, by (iv) of hypothesis 3(c),

$$\begin{aligned}
e_{\min} &\leq e(\eta(C_0)) = e((t_{0,0}, x_0(t_{0,0}), t_{1,0}, x_0(t_{1,0}))) \\
&\leq e((t_{0,0}, X_0^1(t_{0,0}), \dots, X_0^n(t_{0,0}), t_{1,0}, X_0^1(t_{1,0}), \dots, X_0^n(t_{1,0}))) \\
&= \lim_{j \rightarrow \infty} e((t_{0,k_j}, X_{k_j}^1(t_{0,k_j}), \dots, t_{1,k_j}, X_{k_j}^1(t_{1,k_j}), \dots, X_{k_j}^n(t_{1,k_j}))) \\
&= \lim_{j \rightarrow \infty} e(\eta(C_{k_j})) = e_{\min}.
\end{aligned}$$

The theorem is now proved.

REFERENCES

- [1] E. J. McSHANE, *Relaxed controls and variational problems*, this Journal, 5 (1967), pp. 438-485.
- [2] H. H. SCHAEFER, *Topological Vector Spaces*, Macmillan, New York, 1966.

A FINITE SET OF $n \times n$ STOCHASTIC MATRICES GENERATING ALL n -DIMENSIONAL PROBABILITY VECTORS WHOSE COORDINATES HAVE FINITE BINARY EXPANSION*

A. PAZ†

Abstract. For any integer n , it is proved that there exists a finite set of stochastic matrices such that any n -dimensional probabilistic vector, the entries of which have finite binary expansion, can be realized as a row in a finite product of those matrices. The matrices are given explicitly and an algorithm is provided for finding the realization of a given vector.

1. Introduction. Rabin [1] mentions a pair of stochastic matrices that were suggested by E. F. Moore:

$$P_0 = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad P_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}.$$

It can be verified that if

$$P_{\delta_1} P_{\delta_2} \cdots P_{\delta_k} = \begin{bmatrix} m & p \\ q & r \end{bmatrix}, \quad \delta_i = 0 \text{ or } 1,$$

then $p = .\delta_k \delta_{k-1} \cdots \delta_1$, where p is written in binary expansion. Thus any 2-dimensional rational probability vector can be realized as the first row in a product of the above matrices in a proper sequence—except for $(0, 1)$, which is the second row of P_1 .

The above matrices considered as the transition matrices of a 2-state probabilistic automaton were used by Rabin [1] and by Paz [2] for proving various properties of probabilistic automata.

The scope of this paper is to generalize the above example to the n -dimensional case, i.e., to prove the existence of a finite set of n -dimensional *stochastic* matrices, such that any n -dimensional rational probability vector (the entries of which have finite binary expansion) can be realized as a row in a finite product of those matrices.

It is hoped that this generalization, besides its pure mathematical interest, will find uses in the theory of n -dimensional probabilistic automata and related topics (e.g., nonhomogeneous finite Markov chains, finite-state communication channels, etc.).

* Received by the editors November 16, 1966, and in revised form February 17, 1967.

† College of Engineering, University of California, Berkeley, California 94720; on leave from Technion-Israel Institute of Technology, Haifa, Israel. This work was supported in part by the United States Office of Naval Research, Information Systems Branch, under Contract N62558-3882 NR 049-130, and by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under Grant AF-AFOSR-639-66.

We shall prove first some preliminary lemmas. In what follows we consider the set of all probability vectors whose components have a finite binary expansion. (This set will be denoted by $V(n)$.)

Let x be some vector in $V(n)$:

$$x = (x_1, \dots, x_n), \quad x_i = .x_{i1}x_{i2} \dots x_{im_i}, \quad x_{ij} = 0 \text{ or } 1.$$

The length of x , denoted by $l(x)$, is defined as

$$(1) \quad l(x) = \max_i \{m_i : x_{im_i} > 0\}.$$

It is assumed throughout that the lengths of the expansions of all components of x are equal to $l(x)$. This is not a restriction, as one can always add zeros to the expansions if necessary.

2. Matrices $A(\beta)$. Let $\beta \in V(n)$ be the probabilistic vector $(\beta_1, \dots, \beta_n)$ such that $\beta_i = .\beta_{i1} \dots \beta_{iq}$ with $q = l(\beta)$. Let t be an integer, $t \geq 1$. The matrix $A(\beta)$ is the $n \times n$ matrix defined as:

$$(2) \quad A(\beta) = \|a_{ij}\| = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \dots & \beta_n \\ \beta_1 - \frac{1}{2} & \beta_2 + \frac{1}{2} & \beta_3 & \dots & \beta_n \\ \beta_1 - \frac{1}{2} & \beta_2 & \beta_3 + \frac{1}{2} & \dots & \beta_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_1 - \frac{1}{2} & \beta_2 & \beta_3 & \dots & \beta_n + \frac{1}{2} \end{bmatrix}.$$

The following properties of these matrices are easily proved:

- (i) The matrix $A(\beta)$ is uniquely defined by β .
- (ii) The matrix $A(\beta)$ is stochastic if and only if $\beta_1 \geq \frac{1}{2}$ (for β is a stochastic vector).
- (iii) Let $\alpha = (\alpha_1, \dots, \alpha_n)$ be a probabilistic vector $\alpha_i = .\alpha_{i1} \dots \alpha_{ip}$, $p = l(\alpha)$, and let $\beta_1 \geq \frac{1}{2}$; then $\alpha \cdot A(\beta) = (\gamma_1, \dots, \gamma_n) = \gamma$, where γ is a probabilistic vector and $\gamma_i = \beta_i + \frac{1}{2}\alpha_i$, $i = 2, 3, \dots, n$. For $\gamma_i = \sum_{j=1}^n \alpha_j a_{ji} = \alpha_1 \beta_i + \alpha_2 \beta_i + \dots + \alpha_i(\beta_i + \frac{1}{2}) + \alpha_{i+1} \beta_i + \dots + \alpha_n \beta_i = \beta_i \sum_{j=1}^n \alpha_j + \frac{1}{2}\alpha_i = \beta_i + \frac{1}{2}\alpha_i$.

3. Properties of expansions. Let $\alpha = (\alpha_1, \dots, \alpha_n)$ be a probabilistic vector $\alpha \in V(n)$, $n \geq 2$; $\alpha_i = .\alpha_{i1} \dots \alpha_{ip}$, $p = l(\alpha)$.

Write down the expansions of α_i , $i \geq 2$, in the form of a list (called the expansion of α):

$$(3) \quad \begin{array}{c} .\alpha_{21} \dots \alpha_{2p} \\ .\alpha_{31} \dots \alpha_{3p} \\ \vdots \\ .\alpha_{n1} \dots \alpha_{np} . \end{array}$$

Define the integer $z(n)$ as:

$$(4) \quad z(n) = \begin{cases} \log_2(n-1) & \text{if } \log_2(n-1) \text{ is an integer,} \\ [\log_2(n-1)] + 1 & \text{otherwise,} \end{cases}$$

i.e., $z(n) = [\log_2(n-1)]^+$, the least integer $\geq \log_2(n-1)$. Assume now that $\sum_{i=2}^n \alpha_i < 1$ (i.e., $\alpha_1 > 0$). In this case, if for some j we have that $\sum_{i=2}^n \alpha_{ij} \geq 2$, then there is a $k < j$ for which $\sum_{i=2}^n \alpha_{ik} = 0$, as the carry of the binary sum of the j th column must be absorbed by some zero columns ($\sum \alpha_i < 1$).

LEMMA 1. *Let α be a vector, and $z(n)$ the number as defined in (4). Then $z(n)$ consecutive columns of zeros in the expansion of α absorb any possible carry arising from the subsequent columns.*

Proof. Consider the $n-1$ real numbers c_i , $2 \leq i \leq n$:

$$\begin{array}{lll} c_2 = \overbrace{.000 \cdots 0}^x c_{2,k} c_{2,k+1} \cdots & \leq \overbrace{.000 \cdots 0111 \cdots}^x & \leq 2^{-x}, \\ c_3 = .000 \cdots 0 c_{3,k} c_{3,k+1} \cdots & \leq .000 \cdots 0111 \cdots & \leq 2^{-x}, \\ \vdots & & \vdots \\ c_n = .000 \cdots 0 c_{n,k} c_{n,k+1} \cdots & \leq .000 \cdots 0111 \cdots & \leq 2^{-x}. \end{array}$$

The x columns of zeros can absorb all the carrying from the subsequent columns if

$$(n-1) \cdot 2^{-x} \leq .111 \cdots = 1,$$

i.e., if $2^x \geq n-1$ or $x \geq \log_2(n-1)$. Therefore taking $x = z(n)$ is sufficient. (The argument is unchanged if the x columns of zeros are not the first x columns.)

4. The sets $\mathfrak{B}(n)$ and $\mathfrak{A}(n)$. The set $\mathfrak{B}(n)$ is defined as:

$$(5) \quad \mathfrak{B}(n) = \{\beta \mid \beta \in V(n), \beta_1 \geq \tfrac{1}{2}, l(\beta) = k = 0, 1, 2, \dots, z(n) + 1\},$$

where β_1 is the first coordinate of β , $l(\beta)$ is the length of β (see (1)) and $z(n)$ is as in (4).

The set $\mathfrak{A}(n)$ is defined as:

$$(6) \quad \mathfrak{A}(n) = \{A(\beta) \mid \beta \in \mathfrak{B}(n)\},$$

where $A(\beta)$ is as in (2) and $A(\beta)$ is stochastic for $\beta_1 \geq 1/2$ (see property (ii) of the matrices $A(\beta)$). Note that the set $\mathfrak{B}(n)$ (hence also $\mathfrak{A}(n)$) is finite, as the number of n -dimensional rational vectors of length $\leq z(n) + 1$ does not exceed $2^{n(z(n)+1)}$.

The set $\overline{\mathfrak{B}(n)}$ is defined as:

$$(7) \quad \overline{\mathfrak{B}(n)} = \{\gamma \mid \gamma = (\tfrac{1}{2}, \gamma_2, \gamma_3, \dots, \gamma_{n+1}), \\ (\tfrac{1}{2} + \gamma_2, \gamma_3, \dots, \gamma_{n+1}) \in \mathfrak{B}(n)\}.$$

The set $\overline{\mathfrak{A}(n)}$ is defined as:

$$\overline{\mathfrak{A}(n)} = \{A(\beta) \mid \beta \in \overline{\mathfrak{B}(n)}\}.$$

LEMMA 2. $\overline{\mathfrak{B}(n)} \subset \mathfrak{B}(n+1)$, therefore $\overline{\mathfrak{A}(n)} \subset \mathfrak{A}(n+1)$.

Proof. Let $\beta = (1/2, \beta_2, \dots, \beta_{n+1}) \in \overline{\mathfrak{B}(n)}$; then $\beta_1 = 1/2$, $\sum_{i=1}^{n+1} \beta_i = 1$ (by definition, $(1/2 + \beta_2, \beta_3, \dots, \beta_{n+1}) \in \mathfrak{B}(n)$) and $l(\beta) \leq z(n) + 1 \leq z(n+1) + 1$. This implies that $\beta \in \mathfrak{B}(n+1)$.

LEMMA 3. Let $A(\beta)$ be some matrix in $\mathfrak{A}(n)$. There is a matrix $A(\gamma)$ in $\overline{\mathfrak{A}(n)}$ (therefore $A(\gamma)$ is in $\mathfrak{A}(n+1)$) such that

$$A(\gamma) = \begin{bmatrix} \frac{1}{2} & \cdots & \gamma_{n+1} \\ 0 & & A(\beta) \\ \vdots & & \\ 0 & & \end{bmatrix},$$

where γ is a vector of the form:

$$\gamma = (\tfrac{1}{2}, \gamma_2, \dots, \gamma_{n+1}).$$

Proof. Let $\beta = (\beta_1, \dots, \beta_n) \in \mathfrak{B}(n)$; then $\beta_1 \geq 1/2$, so that $\beta_1 = 1/2 + \gamma_2$ for some $\gamma_2 \geq 0$. Set $\gamma = (1/2, \gamma_2, \beta_2, \dots, \beta_n)$. Clearly $\gamma \in \overline{\mathfrak{B}(n)}$, so that $A(\gamma) \in \overline{\mathfrak{A}(n)}$. Consider now the matrix $A(\gamma)$. All entries in the first column of $A(\gamma)$, except the first, are $\gamma_1 - 1/2 = 1/2 - 1/2 = 0$. It is easily verified that all other entries of $A(\gamma)$ are as in the lemma (see the definition of $A(\gamma)$ given in (2)).

COROLLARY. Let $A(\beta^1), \dots, A(\beta^t)$ be matrices in $\mathfrak{A}(n)$, and $A(\gamma^1), \dots, A(\gamma^t)$ be the corresponding matrices in $\mathfrak{A}(n+1)$; then

$$A(\gamma^1)A(\gamma^2) \cdots A(\gamma^t) = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{n+1} \\ 0 & & & \\ & A(\beta^1) & \cdots & A(\beta^t) \\ 0 & & & \end{bmatrix},$$

where $(\alpha_1, \dots, \alpha_{n+1})$ is some probabilistic vector.

5. Examples. Let α be the 3-dimensional vector

$$\alpha = (.0010100011, .1001101011, .0011110010);$$

then $l(\alpha) = 10$, and the expansion of α is

$$.1001101011$$

$$.0011110010.$$

For the above vector, $n = 3$ with $z(3) = [\log_2 2] = 1$. The maximal carry for this vector can be absorbed in a single column of zeros.

The following examples illustrate the definitions and lemmas of the §4.

Example 1. For $n = 2$, $z(n) = [\log_2 1] = 0$:

$$\mathfrak{Q}(2) = \{(\frac{1}{2}, \frac{1}{2}), (1, 0)\};$$

$\mathfrak{Q}(2)$ contains only two matrices, suggested by E. F. Moore (see Introduction).

Example 2. For $n = 3$, $z(n) = [\log_2 2] = 1$:

$\mathfrak{Q}(3)$ contains 6 vectors, namely,

$$\beta^1 = (1, 0, 0) \quad \text{with} \quad l(\beta^1) = 0,$$

$$\beta^2 = (\frac{1}{2}, \frac{1}{2}, 0), \quad \beta^3 = (\frac{1}{2}, 0, \frac{1}{2}) \quad \text{with} \quad l(\beta^2) = l(\beta^3) = 1,$$

$$\beta^4 = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}), \quad \beta^5 = (\frac{3}{4}, \frac{1}{4}, 0), \quad \beta^6 = (\frac{3}{4}, 0, \frac{1}{4})$$

$$\text{with} \quad l(\beta^4) = l(\beta^5) = l(\beta^6) = 2 = z(3) + 1.$$

The matrices in $\mathfrak{Q}(3)$ are:

$$\begin{aligned} A(\beta^1) &= \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, & A(\beta^2) &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, & A(\beta^3) &= \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}, \\ A(\beta^4) &= \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{3}{4} \end{bmatrix}, & A(\beta^5) &= \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}, & A(\beta^6) &= \begin{bmatrix} \frac{3}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{3}{4} \end{bmatrix}. \end{aligned}$$

The sets $\overline{\mathfrak{Q}(2)}$ and $\overline{\mathfrak{Q}(3)}$ are

$$\overline{\mathfrak{Q}(2)} = \{(\frac{1}{2}, 0, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}, 0)\},$$

$$\overline{\mathfrak{Q}(3)} = \{A(\beta^3), A(\beta^2)\}.$$

It is seen that the matrices $A(\beta^2)$ and $A(\beta^3)$ are of the form specified in Lemma 3, and those in $\mathfrak{Q}(2)$ are submatrices of the above.

6. Main theorem. We are now ready to prove the following theorem.

THEOREM. Let n be an integer, $\alpha = (\alpha_1, \dots, \alpha_n)$ any vector in $V(n)$, and $\mathfrak{Q}(n) = \{A_i\}$ the finite set of matrices defined in (6). There exist t_1, \dots, t_m such that α is some row of $A_{t_1} A_{t_2} \dots A_{t_m}$. If $\alpha_1 = \dots = \alpha_k = 0$, then α is the $(k+1)$ st row, and if $\alpha_1 \neq 0$ then α is the first row in the above product.

Proof. For $n = 1$ the theorem is trivially true, as in this case the single vector in $V(1)$ is 1 and the single matrix in $\mathfrak{Q}(1)$ is 1. Assuming now that the theorem holds for $n - 1$, we shall show it to hold for n . Let

$\alpha = (\alpha_1, \dots, \alpha_n)$ be any vector in $V(n)$. If $\alpha_1 = 0$ then the vector $(\alpha_2, \dots, \alpha_n)$ is a vector in $V(n-1)$, and by induction there exist t_1, \dots, t_m such that the vector $(\alpha_2, \dots, \alpha_n)$ is some row in a matrix of the form $A_{t_1} A_{t_2} \cdots A_{t_m}$, $A_t \in \mathcal{A}(n-1)$.

Using the corollary of Lemma 3, we have that $(0, \alpha_2, \dots, \alpha_n)$ is some row in a product matrix $A'_{t_1} A'_{t_2} \cdots A'_{t_m} \in \mathcal{A}(n)$. If $\alpha_2 \neq 0$, then α is the first row in the product of the A_{t_i} , and therefore the second row in the product of the A'_{t_i} . If $\alpha_2 = \cdots = \alpha_k = 0$, then α is the k th row in the product of the A 's, and therefore the $(k+1)$ st row in the product of the A' 's. It remains to be shown that the theorem holds for α with $\alpha_1 > 0$, and this will be proved by induction on $l(\alpha)$.

Case 1. $l(\alpha) = 0$ or $l(\alpha) = 1$. In this case, $\alpha_1 > 0$ implies that $\alpha_1 = 1/2$ or $\alpha_1 = 1$. Thus $\alpha \in \mathcal{B}(n)$ and α is the first row of some matrix in $\mathcal{A}(n)$.

Case 2. $l(\alpha) = m$ for some $m \geq 2$. Two subcases must be considered.

Case 2.1. $0 < \alpha_1 < 1/2$. Then, $\sum_{i=2}^n \alpha_i > 1/2$.

Write down the expansion of α :

$$\begin{array}{cccc} \alpha_{21} & \cdots & \alpha_{2m} & \\ & & & \\ \alpha_{31} & \cdots & \alpha_{3m} & \\ \vdots & & \vdots & \\ \alpha_{n1} & \cdots & \alpha_{nm} & \end{array}$$

and define the integer j as follows: if

$$(9) \quad \sum_{i=2}^n \alpha_{im} 2^{-m} \geq \frac{1}{2}, \quad \text{then } j = m,$$

and j satisfies the following inequalities otherwise:

$$\sum_{k=j}^m \sum_{i=2}^n \alpha_{ik} 2^{-k} \geq \frac{1}{2}, \quad \sum_{k=j+1}^m \sum_{i=2}^n \alpha_{ik} 2^{-k} < \frac{1}{2}.$$

By assumption, $\sum_{i=2}^n \alpha_i > 1/2$, and therefore j is a well-defined integer. Moreover, $j \leq z(n) + 1$ (equality is possible); this follows from the fact that $\sum_{k=j}^m \sum_{i=2}^n \alpha_{ik} 2^{-k} \geq 1/2$ and from Lemma 1. Let l be the maximal integer satisfying

$$(10) \quad \sum_{k=j+1}^m \sum_{i=2}^n \alpha_{ik} 2^{-k} + \sum_{i=2}^l \alpha_{ij} 2^{-j} < \frac{1}{2}.$$

If there is no integer satisfying the above inequality, then $l = 1$. Consider the sum

$$(11) \quad \sum_{k=1}^{j-1} \sum_{i=2}^n \alpha_{ik} 2^{-k} + \sum_{i=l+1}^n \alpha_{ij} 2^{-j}.$$

It follows from the definition of l that the sum (11) is $\leq 1/2$. To prove this,

observe that the sum in (10) is not smaller than $1/2 - (1/2)^j$ (the j th column in the expansion of α has nonzero entries by the definition of j); if the sum in (11) is $> 1/2$, then it is $\geq 1/2 + (1/2)^j$ (for the same reason), and this would imply that the two sums combined are ≥ 1 , contrary to the assumption that the sum of the two sums (which equals the sum $\sum_{i=2}^n \alpha_i$) is smaller than 1. We define now the vectors α^1 and α^2 as follows:

$$\alpha^1 = (\alpha_1^1, \alpha_2^1, \dots, \alpha_n^1)$$

with

$$(12) \quad \begin{aligned} \alpha_i^1 &= \sum_{k=j}^m \alpha_{ik} 2^{-k} & \text{if } 2 \leq i \leq l, \\ \alpha_i^1 &= \sum_{k=j+1}^m \alpha_{ik} 2^{-k} & \text{if } l < i \leq n, \\ \alpha_1^1 &= \frac{1}{2} - \sum_{i=2}^n \alpha_i^1; \end{aligned}$$

and

$$\alpha^2 = (\alpha_1^2, \dots, \alpha_n^2)$$

with

$$(13) \quad \begin{aligned} \alpha_i^2 &= \sum_{k=1}^{j-1} \alpha_{ik} 2^{-k} & \text{if } 2 \leq i \leq l, \\ \alpha_i^2 &= \sum_{k=1}^j \alpha_{ik} 2^{-k} & \text{if } l < i \leq n, \\ \alpha_1^2 &= 1 - \sum_{i=2}^n \alpha_i^2. \end{aligned}$$

It is easily verified that

$$\alpha_i = \alpha_i^1 + \alpha_i^2 \quad \text{for } i = 2, 3, \dots, n.$$

Consider the vector $2\alpha^1 = \beta$. By definition, we have that $\beta \in V(n)$ and $l(\beta) = m - 1$. By induction, β can be realized as the first row in a product of matrices:

$$A_{t_1} A_{t_2} \cdots A_{t_r}, \quad A_{t_i} \in \mathfrak{Q}(n).$$

By the definition of α^2 , we have that $\alpha^2 \in V(n)$, $l(\alpha^2) \leq z(n) + 1$ (as remarked before, j in (9) is $\leq z(n) + 1$) and $\alpha_1^2 \geq 1/2$ (for the sum (11), which equals $\sum_{i=2}^n \alpha_i^2$, is $\leq 1/2$). Thus, $\alpha^2 \in \mathfrak{B}(n)$.

It follows immediately that α is the first row in the product matrix

$$A_{t_1} A_{t_2} \cdots A_{t_r} \cdot A(\alpha^2), \quad A(\alpha^2) \in \mathfrak{Q}(n),$$

as (using property (iii) of the matrices $A(\beta)$) the i th entry in the first row of the above product equals

$$\frac{1}{2}(2\alpha_i^1) + \alpha_i^2 = \alpha_i, \quad i = 2, 3, \dots, n,$$

as required.

Case 2.2. $\frac{1}{2} \leq \alpha_1 \leq 1$. Hence, $0 \leq \sum_{i=2}^m \alpha_i \leq 1/2$.

Consider the vector $\beta = (\beta_1, \dots, \beta_n)$ with

$$\beta_i = 2\alpha_i, \quad i = 2, 3, \dots, n,$$

$$\beta_1 = 1 - \sum_{i=2}^n \beta_i.$$

By assumption $l(\alpha) > 1$ so $\alpha_1 < 1$, i.e., $\beta_1 > 0$. It follows from the definition of β that $l(\beta) = m - 1$ and, by induction, β can be realized as the first row in a product of matrices:

$$A_{t_1} A_{t_2} \cdots A_{t_r}, \quad A_{t_i} \in \mathfrak{A}(n).$$

Let γ be defined as $\gamma = (\gamma_1, \dots, \gamma_n)$ with $\gamma_1 = 1$, $\gamma_i = 0$, $i = 2, 3, \dots, n$; then $\gamma \in \mathfrak{B}(n)$, and α is the first row in the product matrix

$$A_{t_1} A_{t_2} \cdots A_{t_r} \cdot A(\gamma), \quad A(\gamma) \in \mathfrak{A}(n).$$

(This follows immediately from property (iii) of the matrices $A(\beta)$.) The proof is thus complete.

7. Algorithm. The proof of our main theorem is constructive and, in fact, provides an algorithm for generating any n -dimensional vector in $V(n)$ from the matrices in $\mathfrak{A}(n)$. The algorithm, based on a kind of "push-down" technique, closely resembles the well-known Euclidean algorithm for converting ordinary into binary fractions. It will be illustrated by the following examples.

Let α be the 6-dimensional vector

$$\alpha = \left(\frac{21}{64}, \frac{15}{64}, \frac{7}{64}, \frac{7}{64}, \frac{7}{64}, \frac{7}{64}\right).$$

The expansion of α is:

$$\begin{array}{r} .001111 \\ .000111 \\ .000111 \\ .000111 \\ .000111 \\ .000111. \end{array}$$

The integer j , defined in (9), is $j = 4$, and l , defined in (10), is $l = 5$. The vectors α^1 and α^2 , defined in (12) and (13), are, therefore,

$$\alpha^2 = (\frac{1}{16}, \frac{1}{8}, 0, 0, 0, \frac{1}{16}) \quad \text{with} \quad \alpha^2 \in \mathfrak{B}(6),$$

$$\alpha^1 = (\frac{1}{64}, \frac{7}{64}, \frac{7}{64}, \frac{7}{64}, \frac{7}{64}, \frac{3}{64}).$$

Proceeding analogously with $2\alpha^1$ (which is in $V(n)$) instead of α , we have

$$2\alpha^1 = (\frac{1}{32}, \frac{7}{32}, \frac{7}{32}, \frac{7}{32}, \frac{7}{32}, \frac{3}{32}) = \beta$$

with

$$\beta^2 = (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, 0) \quad \text{with} \quad \beta^2 \in \mathfrak{B}(6),$$

$$\beta^1 = (\frac{1}{32}, \frac{3}{32}, \frac{3}{32}, \frac{3}{32}, \frac{3}{32}, \frac{3}{32}).$$

Repeating for $2\beta^1 = \gamma \in V(n)$ we have

$$\gamma = (\frac{1}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}),$$

$$\gamma^2 = (\frac{1}{2}, 0, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}) \quad \text{with} \quad \gamma^2 \in \mathfrak{B}(6),$$

$$\gamma^1 = (\frac{1}{16}, \frac{3}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}).$$

Then $2\gamma^1 = \xi \in V(n)$, which yields

$$\xi = (\frac{1}{8}, \frac{3}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}),$$

$$\xi^2 = (\frac{1}{2}, \frac{1}{4}, 0, 0, \frac{1}{8}, \frac{1}{8}) \quad \text{with} \quad \xi^2 \in \mathfrak{B}(6),$$

$$\xi^1 = (\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, 0, 0),$$

and $2\xi^1 = \eta \in V(n)$; hence,

$$\eta = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, 0, 0),$$

$$\eta^2 = (\frac{1}{2}, 0, \frac{1}{4}, \frac{1}{4}, 0, 0) \quad \text{with} \quad \eta^2 \in \mathfrak{B}(6),$$

$$\eta^1 = (\frac{1}{4}, \frac{1}{4}, 0, 0, 0, 0).$$

Finally,

$$2\eta^1 = (\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0) = \rho$$

is in $\mathfrak{B}(6)$. We thus have that α is the first row in the following product of matrices:

$$A(\rho) \cdot A(\eta^2) \cdot A(\xi^2) \cdot A(\gamma^2) \cdot A(\beta^2) \cdot A(\alpha^2),$$

and by the properties of these matrices we have that

$$(14) \quad \alpha_i = 2^{-5}\rho_i + 2^{-4}\eta_i^2 + 2^{-3}\xi_i^2 + 2^{-2}\gamma_i^2 + 2^{-1}\beta_i^2 + \alpha_i^2,$$

$$i = 2, 3, \dots, n.$$

There is close resemblance between (14) and the usual binary expansion of ordinary fractions. In the light of the above example it would be natural

to extend the discussion to vectors having infinite expansions, but this is not attempted here. As a final remark, note that although any vector in $V(n)$ can be generated as some row in a product of matrices from $\mathcal{A}(n)$ the generation is not unique. For example, for the 3-dimensional case (see §5) we have that the first row in $A(\beta^2) \cdot A(\beta^1)$ equals its counterpart in the matrix $A(\beta^4)$. On the other hand, it seems that the number of matrices in $\mathcal{A}(n)$ is minimal in the sense that a set of matrices containing fewer matrices would not suffice for generating all vectors in $V(n)$.

REFERENCES

- [1] M. O. RABIN, *Probabilistic automata*, Information and Control, 6 (1963), pp. 230–245. (Also in *Sequential Machines: Selected Papers*, E. F. Moore, ed., Addison-Wesley, Reading, Massachusetts, 1964, pp. 98–114.)
- [2] A. PAZ, *Some aspects of probabilistic automata*, Ibid., 9 (1966), pp. 26–60.

ON FUNCTION SPACE PURSUIT-EVASION GAMES*

WILLIAM A. PORTER†

Summary. Functional analysis is used to analyze a minimax problem for a class of functionals on a Hilbert space. The abstract solution is then interpreted, through examples, in the context of pursuit-evasion games. The examples given are chosen to emphasize the breadth of the analysis. Some computational aspects of the minimax problem are considered.

1. Introduction. Problems in optimal control can frequently be formulated in terms of the minimization or maximization of a functional on a Banach space. Moreover, "two-sided" optimization problems, such as pursuit-evasion games, can often be identified with saddle points of Banach space functionals. For instance, suppose that u and v denote the inputs to a pursuit system and an evasion system, respectively. Let f denote a functional on the tuple (u, v) such that the pursuer attempts to minimize f while the evader attempts to maximize f . A tuple (u_0, v_0) such that

$$f(u, v_0) \geq f(u_0, v_0) \geq f(u_0, v)$$

for all u and v is optimal from both the pursuit and the evasion points of view.

The present paper has two main objectives. First, the analysis of pursuit-evasion games. This class of problems is an abstraction of the problem considered by Ho, Bryson and Barron in [1]. Secondly, attention is focused on the use of function space methods and the generality of results obtained by such an approach. Thus Hilbert spaces provide the setting, and the method of analysis follows closely the approach of Rall in [2].

2. Basic notions. To proceed with reasonable efficiency it is necessary to assume a working knowledge of Hilbert spaces (see [3], [4] or [6]). This section, of necessity, is limited to introducing the minimal notation that is used throughout the article.

Consider two real Hilbert spaces H_1 and H_2 , with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ respectively. Let $H = H_1 \times H_2$ denote the usual Cartesian product of H_1 and H_2 ,

$$H = \{ \theta = (\theta_1, \theta_2) : \theta_1 \in H_1, \theta_2 \in H_2 \}.$$

H is also a Hilbert space when equipped with the inner product

* Received by the editors September 14, 1966, and in revised form February 20, 1967.

† Department of Electrical Engineering, University of Michigan, Ann Arbor, Michigan 48104. The research for this paper was supported in part by the United States Army Research Office, Durham, under Contract DA-31-124-ARD-D-391.

$$\langle \theta, \psi \rangle = \langle \theta_1, \psi_1 \rangle_1 + \langle \theta_2, \psi_2 \rangle_2, \quad \theta, \psi \in H.$$

Suppose now that $f: H \rightarrow R$ is a real-valued functional on H which is twice Fréchet differentiable (see [4], [5] or [9]). The first derivative, $f'(\theta)$, of f at $\theta = (\theta_1, \theta_2)$ is a linear functional on H and thus has the form

$$\langle f'(\theta), \psi \rangle = \langle f'_1(\theta), \psi_1 \rangle_1 + \langle f'_2(\theta), \psi_2 \rangle_2,$$

where the partial derivatives $f'_1(\theta)$ and $f'_2(\theta)$ are formed using the rules for differentiation in a product space. With the meaning of this expression we write $f'(\theta) = (f'_1(\theta), f'_2(\theta))$. The element $\theta \in H$ is called a critical point of f if $f'(\theta) = 0$. If θ_0 is a point of local maximum or local minimum for f (that is, $f(\theta) \leq f(\theta_0)$ or $f(\theta) \geq f(\theta_0)$ in some ball about θ_0), then θ_0 is necessarily a critical point of f . In this paper, however, our interest lies primarily with the saddle points of f .

Consider the functionals $f_1: H_1 \rightarrow R$ and $f_2: H_2 \rightarrow R$ defined by the expressions

$$(1) \quad f_1(u) = f(u, v), \quad u \in H_1,$$

$$(2) \quad f_2(v) = f(u, v), \quad v \in H_2.$$

Thinking of v as a parameter, u is to minimize f_1 . Conversely, thinking of u as a parameter, v is to maximize f_2 . A solution tuplet (u_0, v_0) is also of necessity a critical point of f and, in this case, has the property

$$f(u_0, v_0) = \min_u \max_v f(u, v) = \max_v \min_u f(u, v)$$

and for transparent reasons is referred to as a *minimax* point of f .

3. A specific class of problems. Let $T: H_1 \rightarrow H_3$ and $S: H_2 \rightarrow H_3$ denote bounded linear transformations and $\xi \in H_3$ a fixed element. Let f be the real-valued functional on H defined by

$$(3) \quad f(u, v) = \frac{1}{2} \{ \| Sv - Tu + \xi \|^2 + \| u \|^2 - \| v \|^2 \}, \quad (u, v) \in H.$$

Our immediate objective is to locate the minimax points of f . An important step in this direction is given by the following lemma.

LEMMA 1. For f defined in (3),

$$f'_1(u, v) = u - T^* \xi - T^* Sv + T^* Tu,$$

$$f'_2(u, v) = -v + S^* \xi - S^* Tu + S^* Sv.$$

Proof. The proof is straightforward and will only be sketched here. First, expand f in terms of inner products, for instance,

$$\begin{aligned} \| Sv + \xi - Tu \|^2 &= \langle u, T^* Tu \rangle + \langle v, SS^* v \rangle + \langle \xi, \xi \rangle + 2 \langle v, S^* \xi \rangle \\ &\quad - 2 \langle Sv, Tu \rangle - 2 \langle T^* \xi, u \rangle. \end{aligned}$$

The right-hand side of this expression is easily differentiated term by term. For example, from the limit

$$\begin{aligned} \lim_{\delta u \rightarrow 0} & \left\{ \frac{\langle u + \delta u, T^*T^*(u + \delta u) \rangle - \langle u, T^*Tu \rangle}{\|\delta u\|} \right\} \\ &= \lim_{\delta u \rightarrow 0} \left\{ \frac{2 \langle T^*Tu, \delta u \rangle + \langle \delta u, T^*T\delta u \rangle}{\|\delta u\|} \right\} \\ &= \lim_{\delta u \rightarrow 0} \left\{ 2 \left\langle T^*Tu, \frac{\delta u}{\|\delta u\|} \right\rangle \right\}, \end{aligned}$$

we see that

$$\frac{\partial}{\partial u} \langle u, T^*Tu \rangle = 2T^*Tu.$$

Repeating this process term by term and collecting results leads directly to the assertion of the lemma.

The critical points of f must of necessity satisfy the conditions $f_1' = 0$ and $f_2' = 0$. Using the lemma, we arrive immediately at the conditions

$$\begin{aligned} u &= T^*\xi + T^*Sv - T^*Tu, \\ v &= S^*\xi + S^*Sv - S^*Tu. \end{aligned}$$

Notice that the first equation requires u to be in the range of T^* while the second equation places v in the range of S^* . That is, for some $\lambda, \beta \in H_3$,

$$u = T^*\lambda, \quad v = S^*\beta.$$

Making these changes of variables results in

$$\begin{aligned} T^*\lambda &= T^*(\xi + SS^*\beta - TT^*\lambda), \\ S^*\beta &= S^*(\xi + SS^*\beta - TT^*\lambda), \end{aligned}$$

which will hold whenever

$$\begin{aligned} (4) \quad \lambda &= \xi + SS^*\beta - TT^*\lambda, \\ \beta &= \xi + SS^*\beta - TT^*\lambda, \end{aligned}$$

from which it follows immediately that $\lambda = \beta$ and, moreover,

$$(5) \quad [I + TT^* - SS^*]\lambda = \xi.$$

Thus, whenever the indicated inverse exists, the critical points (u_0, v_0) of f are of the form

$$(6) \quad u_0 = T^*[I + TT^* - SS^*]^{-1}\xi,$$

$$(7) \quad v_0 = S^*[I + TT^* - SS^*]^{-1}\xi.$$

These results are summarized in the following corollary to Lemma 1.

COROLLARY. If (u_0, v_0) is a critical point for f of (3), then $u_0 = T^*\lambda$ and $v_0 = S^*\lambda$, where λ is some solution of (5). Where the indicated inverse exists, u_0 and v_0 are determined uniquely by (6) and (7) respectively.

Remark 1. In [1] a functional of the form

$$(8) \quad \frac{1}{2} \{ \| Sv - Tu + \xi \|^2 + a^2 \| u \|^2 - b^2 \| v \|^2 \}$$

was considered where $a, b > 0$ are parameters. Other possible variations on (3) include

$$(9) \quad \frac{1}{2} \{ \| Sv - Tu + \xi \|^2 + \| u - u_0 \|^2 - \| v - v_0 \|^2 \},$$

$$(10) \quad \frac{1}{2} \{ \| Sv - Tu + \xi \|^2 + \| u \|^2 - \| v \|^2 + 2 \langle u, Kv \rangle \},$$

where $u_0 \in H_1$, $v_0 \in H_2$ are fixed elements, and $K: H_2 \rightarrow H_1$ is linear. Each of these and/or combinations of these variations reduce to $f(u, v)$ by an elementary change of variables. In (8) it suffices to set $u' = au$, $v' = bv$, $S' = (1/b)S$, and $T' = (1/a)T$. In (9) set $u' = u - u_0$, $v' = v - v_0$, $\xi' = \xi + Tu_0 - Sv_0$. In (10) an appropriate change of variables is $u' = u + Kv$, $S' = S + TK$, while H_2 is equipped with the (equivalent) inner-product $(x, y) = \langle x, (I + K^*K)y \rangle$ for $x, y \in H_2$. The solutions for these cases are also obtained by direct substitution in (6) and (7). Hence no generality is lost in concentrating attention on the functional of (3).

Example 1. In [1] the following pursuit-evasion problem is considered. The pursuit and evasion systems satisfy the multivariate differential equations¹

$$(11) \quad \begin{aligned} \dot{x}_p(t) &= F_p(t)x_p(t) + G_p(t)u(t), & x_p(t_0) &= x_p^0; \\ \dot{x}_e(t) &= F_e(t)x_e(t) + G_e(t)v(t), & x_e(t_0) &= x_e^0, \end{aligned}$$

respectively. The functional

$$J(u, v) = \frac{1}{2} \{ \| x_e(T) - x_p(T) \|^2 + \| u \|^2 - \| v \|^2 \}$$

is to be maximized (minimized) by the evader (pursuer). In short, the problem is to find, if they exist, control pairs (u_0, v_0) which are saddle points of J .

To bring this problem within the framework of the section, the transformations

$$\begin{aligned} Tu &= \int_{t_0}^T \Phi_p(T, s)G_p(s)u(s) ds, & u &\in H_1, \\ Sv &= \int_{t_0}^T \Phi_e(T, s)G_e(s)v(s) ds, & v &\in H_2, \end{aligned}$$

¹ For this example the notation of [1] is adopted. For simplicity (and at no loss in generality) the weighting matrices R_p , R_e and A^*A , used in [1], have been taken as identity matrices. Thus the Hilbert spaces H_1 and H_2 are appropriate finite copies of $L_2(t_0, T)$ equipped with the usual inner product while H_3 is E^n .

are defined, where Φ_p and Φ_e denote transition matrices. Consequently,

$$x_e(T) - x_p(T) = Sv - Tu + \xi,$$

where

$$\xi = \Phi_e(T, t_0)x_e^0 - \Phi_p(T, t_0)x_p^0,$$

and it then follows by inspection that we are dealing with a concrete version of the main problem.

For the function spaces involved, T^* and S^* are identified by the equations (see [3, §3.3, Example 5])

$$(T^*\lambda)(t) = G_p^*(t)\Phi_p^*(T, t)\lambda, \quad \lambda \in R^n,$$

$$(S^*\beta)(t) = G_e^*(t)\Phi_e^*(T, t)\beta, \quad \beta \in R^n.$$

Consequently, $I + TT^* - SS^*$ may be identified as the matrix

$$I + \int_{t_0}^T \Phi_p(T, s)G_p(s)G_p^*(s)\Phi_p^*(T, s) ds - \int_{t_0}^T \Phi_e(T, s)G_e(s)G_e^*(s)\Phi_e^*(T, s) ds,$$

which is the matrix $K(T, t_0)$ of [1]. Similarly, the optimal controls, namely,

$$(12) \quad \begin{aligned} u_0(t) &= G_p^*(t)\Phi_p^*(T, t)[I + TT^* - SS^*]^{-1}\xi, \quad t \in [t_0, T], \\ v_0(t) &= G_e^*(t)\Phi_e^*(T, t)[I + TT^* - SS^*]^{-1}\xi, \quad t \in [t_0, T], \end{aligned}$$

agree with those computed in [1].

Remark 2. From the construction of the first example it is apparent that a multitude of variations might be easily considered. When the matrices R_p and R_e of [1] are present, the Hilbert spaces change slightly with corresponding changes in the computations of T^* and S^* (see [3, §4.4, Examples 1 and 2]). Similarly, classes of sample data and composite systems fall within the domain of the analysis. The interested reader is referred to [3, §2.5] for further discussion of such applications.

Example 2. For contrast, the minimax problem of this section will now be interpreted in the concrete setting of distributive systems. Since any comprehensive treatment would involve quite lengthy discussions, it is necessary to ration ourselves here to a sketch of the situation. The reader is referred to [11], [12] and [3, Appendix 9] for more complete discussions which include several physical examples.

Consider two distributive systems, defined on the spatial domain $(0, b)$, the time domain (t_0, t_f) , and satisfying the diffusion equations

$$(13) \quad \begin{aligned} x_t(t, \alpha) &= [k(t, \alpha)x_\alpha(t, \alpha)]_\alpha + u(t, \alpha), \quad (t, \alpha) \in (t_0, t_f) \times (0, b), \\ y_t(t, \alpha) &= [l(t, \alpha)y_\alpha(t, \alpha)]_\alpha + v(t, \alpha), \quad (t, \alpha) \in (t_0, t_f) \times (0, b). \end{aligned}$$

Here subscripts denote partial derivatives, u and v denote system inputs, x and y denote system responses, while k and l are diffusion functions of the two systems. These equations are supplemented with auxiliary (i.e., initial and boundary) conditions sufficient to make the problem well defined. The minimax functional of (3) takes the form²

$$f(u, v) = \int_{t_0}^{t_f} \int_0^b \{ [y(t, \alpha) - x(t, \alpha)]^2 + u^2(t, \alpha) - v^2(t, \alpha) \} d\alpha dt.$$

In many important applications the auxiliary conditions are such that the system equations can be identified (à la Sturm-Liouville) with an infinite family of ordinary differential equations. To illustrate, let k and l be constant. Then for the auxiliary conditions $x(t, 0) = x(t, b) = y(t, 0) = y(t, b) = 0$ and $x(t_0, \alpha) = x^0(\alpha)$, $y(t_0, \alpha) = y^0(\alpha)$, equation set (13) may be identified with the equation set³

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), & x(t_0) &= x^0, & t &\in (t_0, t_f), \\ \dot{y}(t) &= Fy(t) + Gv(t), & y(t_0) &= y^0, & t &\in (t_0, t_f). \end{aligned}$$

Here, x , y , u and v are time-varying elements in l_2 while A and F are the infinite diagonal matrices

$$\begin{aligned} A &= \text{diag} \left[-k \left(\frac{\pi}{b} \right)^2 \cdots -k \left(\frac{n\pi}{b} \right)^2 \cdots \right], \\ F &= \text{diag} \left[-l \left(\frac{\pi}{b} \right)^2 \cdots -l \left(\frac{n\pi}{b} \right)^2 \cdots \right]. \end{aligned}$$

If u and v are free of spatially concentrated forces, then B and G are identity operators on l_2 . The j th component of x is related to the function x by the formula

$$[x(t)]_j = \sqrt{\frac{2}{b}} \int_0^b x(t, \alpha) \sin \left(\frac{j\pi\alpha}{b} \right) d\alpha, \quad t \in [t_0, t_f], \quad j = 1, 2, \dots.$$

The tuplets x^0 , y^0 and y are defined similarly.

Now it can be shown that the concept of state transition matrix (also the matrix exponent) generalizes naturally to infinite dimensions (see [13] in addition to the references cited above). Indeed, all equations in Example 1 remain valid in infinite dimensions. Moreover, the solution (u_0, v) , given by the extended version of (12), can be used to define directly the corresponding optimal scalar-valued controls $u_0(t, \alpha)$, $v_0(t, \alpha)$ for (t, α)

² These equations result from the natural Fourier expansions defined by (13). Details are included in the references cited.

³ The Hilbert spaces H_1, H_2 consist here of all measurable and square integrable functions on the rectangle $(t_0, t_f) \times (0, b)$.

$\in (t_0, t_f) \times (0, b)$. To continue, the matrices $\Phi_p(t_f, t)$ and $\Phi_e(t_f, t)$ are infinite and diagonal:

$$\Phi_p(t_f, t) = \text{diag} \left[\cdots \exp \left\{ -k \left(\frac{n\pi}{b} \right)^2 (t_f - t) \right\} \cdots \right], \quad t \in [t_0, t_f],$$

$$\Phi_e(t_f, t) = \text{diag} \left[\cdots \exp \left\{ -l \left(\frac{n\pi}{b} \right)^2 (t_f - t) \right\} \cdots \right], \quad t \in [t_0, t_f],$$

as is the matrix $I + TT^* - SS^*$, which has the typical j th diagonal element

$$1 + \frac{1 - \exp \{ -2k\mu_j(t_f - t_0) \}}{2k\mu_j} - \frac{1 - \exp \{ -2l\mu_j(t_f - t_0) \}}{2l\mu_j},$$

where $\mu_j = (j\pi/b)^2$, $j = 1, 2, \dots$. This latter infinite matrix has an obvious inverse. The tuple $\xi = (\xi_1, \dots, \xi_n, \dots) \in l_2$ is determined by

$$\xi_i = \exp \{ -l\mu_i(t_f - t_0) \} (x_e^0)_i - \exp \{ -k\mu_i(t_f - t_0) \} (x_p^0)_i,$$

where

$$(x_e^0)_i = \sqrt{\frac{2}{b}} \int_0^b x_e^0(\alpha) \sin \left(\frac{i\pi\alpha}{b} \right) d\alpha, \quad i = 1, 2, \dots,$$

and $(x_p^0)_i$ is defined in the same manner. Putting all the parts together the optimal pursuit control is given by

$$u_0(t, \alpha) = \sqrt{\frac{2}{b}} \sum_{i=1}^{\infty} [I + TT^* - SS^*]^{-1} \xi_i \sin \left(\frac{i\pi\alpha}{b} \right) \cdot \exp \left\{ -k \left(\frac{i\pi}{b} \right)^2 (t_f - t) \right\}.$$

The optimal evasion control differs only in that k becomes l .

Example 3. It is not surprising to find a relationship between the min-max problem of this section and certain problems in electronic counter-measure. In the following this relationship is partially exposed. Because the earlier analysis is basically open loop in character and because we wish to avoid the added complexities of introducing Hilbert spaces of random processes, this example should be viewed as a minimal connection between the two problem areas.

For the present discussion all Hilbert spaces are finite Cartesian products of $L_2(-\infty, \infty)$, denoted simply by L_2 hereafter, equipped with the usual norm. Let $\xi \in H_3$ be thought of as a signal which is to be tracked by the output of a linear system. A second linear system is attempting to disrupt

(i.e., countermeasure) the tracking process. The tracking system is modeled by the linear operator $T: H_1 \rightarrow H_3$ while the countersystem has the model $S: H_2 \rightarrow H_3$. To avoid solutions which call for infinite energy on behalf of both systems a concrete form of the functional (3) is selected for design purposes.

Let F denote the Fourier transform on L_2 defined by

$$\sqrt{2\pi}(Fx)(\omega) = \lim_{n \rightarrow \infty} \int_{-n}^n e^{i\omega t} x(t) dt, \quad \omega \in (-\infty, \infty).$$

A theorem of Plancherel (see [14, p. 51]) establishes that F is a one-to-one, onto, norm preserving mapping (i.e., a congruence) on L_2 . The definition of F generalizes to finite products of L_2 and maintains these properties. For convenience, the convention $\hat{x} = Fx$ will be used. Suppose that the system T is stationary. Then a multidimensional version of a theorem of Bochner (see [15, Chap. IV]) states that T has a multiplicative representation in the sense that there exists a matrix, \hat{T} , of uniformly bounded measurable functions such that $y(t) = (Tx)(t) \leftrightarrow \hat{y}(\omega) = \hat{T}(\omega)\hat{x}(\omega)$ for all $x \in H_1$. For convolutions a nonanticipatory T will have a \hat{T} which is analytic in the upper half of the complex plane.

Consider now the linear term in the expansion of $f(u + \delta u, v + \delta v) - f(u, v)$, namely,

$$\langle (I + T^*T)u - T^*(Sv + \xi), \delta u \rangle + \langle (S^*S - I)v - S^*(Tu - \xi), \delta v \rangle.$$

Using Fourier transforms, this is equivalent to the expression

$$\langle (I + \hat{T}^*\hat{T})\hat{u} - \hat{T}^*(\hat{S}\hat{v} + \hat{\xi}), \hat{\delta u} \rangle + \langle (\hat{S}^*\hat{S} - I)\hat{v} - \hat{S}^*(\hat{T}\hat{u} - \hat{\xi}), \hat{\delta v} \rangle.$$

Suppose that we require all signals, except possibly ξ , to have zero values for $t \leq 0$. Then the first inner product is zero for all such δu whenever the function $(I + \hat{T}^*\hat{T})\hat{u} - \hat{T}^*(\hat{S}\hat{v} + \hat{\xi})$ is analytic in the upper half-plane. Using conventional separation methods (see [16, §16.3] or [10, Chap. 2]) for this Wiener-Hopf-like expression, the optimal choice of u is seen to be

$$\hat{u}(\omega) = \hat{\psi}^{-1}(\omega) \{ \hat{\psi}^{-1}(-\omega) \hat{T}^*(\omega) [\hat{S}^*(\omega) \hat{v}(\omega) + \hat{\xi}(\omega)] \}_{\text{u.h.p.}},$$

$$\omega \in (-\infty, \infty),$$

and, similarly, the optimal choice for v is given by

$$\hat{v}(\omega) = \hat{\Delta}^{-1}(\omega) \{ \hat{\Delta}^{-1}(-\omega) \hat{S}^*(\omega) [\hat{T}(\omega) \hat{u}(\omega) - \hat{\xi}(\omega)] \}_{\text{u.h.p.}},$$

$$\omega \in (-\infty, \infty).$$

In these expressions the subscript u.h.p. denotes the upper half-plane part and the factorizations, $\psi(-\omega)\psi(\omega) = I + \hat{T}^*(\omega)\hat{T}(\omega)$ and $\hat{\Delta}(-\omega) = \hat{S}^*(\omega)\hat{S}(\omega) - I$, are assumed, where $\hat{\psi}(\omega)$ and $\hat{\Delta}(\omega)$ are analytic in

the upper half-plane. The reader is referred to [10, Chap. 2] for a discussion of these problems.

Since this solution involves simultaneous nonlinear equations, another (not necessarily simpler) form of the result will be given. To do so, let the tuples

$$\hat{x} = (\hat{u}, \hat{v}), \quad \delta \hat{x} = (\delta \hat{u}, \delta \hat{v}), \quad \hat{z} = \left(\begin{smallmatrix} \hat{\xi} \\ \hat{\xi} \end{smallmatrix} \right)$$

and the matrices

$$\hat{M}(\omega) = \hat{K}(-\omega)\hat{K}(\omega) = \begin{bmatrix} I + \hat{T}^*(\omega)\hat{T}(\omega) & -\hat{T}^*(\omega)\hat{S}(\omega) \\ -\hat{S}^*(\omega)\hat{T}(\omega) & \hat{S}^*(\omega)\hat{S}(\omega) - I \end{bmatrix},$$

$$\hat{Q}(\omega) = \begin{bmatrix} -\hat{T}^*(\omega) & 0 \\ 0 & \hat{S}^*(\omega) \end{bmatrix}$$

be defined. Then the linear term may be written $\langle \hat{M}\hat{x} - \hat{Q}\hat{z}, \delta \hat{x} \rangle$ with the obvious meaning. By the same reasoning as above, the optimal choice of x is given by

$$\hat{x}(\omega) = \hat{K}^{-1}(\omega) \{ \hat{K}^{-1}(-\omega) \hat{Q}(\omega) \hat{z}(\omega) \}_{\text{u.h.p.}}, \quad \omega \in (-\infty, \infty).$$

Here the problem of simultaneous equations is avoided at the expense of increased difficulty in computing the corresponding inverse matrices.

Example 4. The pursuit-evasion problem considered in Example 1 is just one of several that might be formulated within the context of that setting. As a final example, consider the case where the pursuer and evader are in instantaneous competition along their respective trajectories, the "game functional" being

$$I(u, v) = \frac{1}{2} \{ \|x_e - x_p\|^2 + \|u\|^2 - \|v\|^2 \}.$$

Here H_1 and H_2 are as before but H_3 is now a third finite Cartesian product of $L_2(t_0, T)$.

The transformations T and S are defined by

$$(Tu)(t) = \int_{t_0}^t \Phi_p(t, s) G_p(s) u(s) ds, \quad t \in [t_0, T],$$

$$(Sv)(t) = \int_{t_0}^t \Phi_e(t, s) G_e(s) v(s) ds, \quad t \in [t_0, T],$$

while ξ becomes the time-varying tuple

$$\xi(t) = \Phi_e(t, t_0) x_e^0 - \Phi_p(t, t_0) x_p^0, \quad t \in [t_0, T].$$

The adjoints of T and S differ also from the first example, being of the form

$$(T^*\lambda)(t) = \int_t^T G_p^*(t)\Phi_p^*(s, t)\lambda(s) ds, \quad t \in [t_0, T],$$

$$(S^*\beta)(t) = \int_t^T G_e^*(t)\Phi_e^*(s, t)\beta(s) ds, \quad t \in [t_0, T].$$

Using these expressions the operator $I + TT^* - SS^*$ is evidently well defined as is the function λ and the critical point (u_0, v_0) .

It is evident from these expressions that the relationship $\xi \rightarrow \lambda$ is, in this case, an integral equation rather than a matrix equation. This relationship can also be expressed in differential equation form. To do so we make use of the variables $u = T^*\lambda$ and $v = S^*\lambda$, and then we write $[I + TT^* - SS^*]\lambda = \xi$ in the form

$$\lambda = Sv - Tu + \xi.$$

Using the definition of ξ and the transformations T and S it follows from inspection that

$$\lambda(t) = x_e(t) - x_p(t),$$

where

$$\begin{aligned} \dot{x}_p(t) &= F_p(t)x_p(t) + G_p(t)u(t), & x_p(t_0) &= x_p^0, \\ \dot{x}_e(t) &= F_e(t)x_e(t) + G_e(t)v(t), & x_e(t_0) &= x_e^0. \end{aligned}$$

Now using the defining equation for T^* , we have

$$\begin{aligned} u(t) &= (T^*\lambda)(t) = \int_t^T G_p^*(t)\Phi_p^*(s, t)\lambda(s) ds \\ &= G_p^*(t) \left\{ \Phi_p^*(t_0, t) \int_{t_0}^T \Phi_p^*(s, t_0)\lambda(s) ds - \int_{t_0}^t \Phi_p^*(s, t)\lambda(s) ds \right\}, \end{aligned}$$

which is recognized as the integral form of the equation set

$$\begin{aligned} \dot{p}(t) &= -F_p^*(t)p(t) - \lambda(t), \\ p(t_0) &= \int_{t_0}^T \Phi_p^*(s, t_0)\lambda(s) ds, \\ u(t) &= G_p^*(t)p(t). \end{aligned}$$

Similar results hold for $v = S^*\lambda$, and altogether the resulting equations are

$$\begin{aligned} \dot{x}_p(t) &= F_p(t)x_p(t) + G_p(t)G_p^*(t)p(t), \\ \dot{x}_e(t) &= F_e(t)x_e(t) + G_e(t)G_e^*(t)q(t), \\ \dot{p}(t) &= x_p(t) - x_e(t) - F_p^*(t)p(t), \\ \dot{q}(t) &= x_p(t) - x_e(t) - F_e^*(t)q(t). \end{aligned} \tag{14}$$

This homogeneous equation set together with the initial conditions x_p^0 , x_e^0 and the relationships

$$(15) \quad \begin{aligned} p(t_0) &= \int_{t_0}^T \Phi_p^*(s, t_0)[x_e(s) - x_p(s)] ds, \\ q(t_0) &= \int_{t_0}^T \Phi_e^*(s, t_0)[x_e(s) - x_p(s)] ds \end{aligned}$$

complete the description of the problem.

Remark 4. Solving explicitly for $p(t_0)$ in Example 4 is the same sort of two-point boundary value problem that frequently occurs in minimum energy control problems. Once the transition matrix for (14) is computed the interrelationships between $p(t_0)$, $q(t_0)$, $x_p(t_0)$ may be determined from (15). It also seems reasonable to anticipate a closed loop minimax controller. Assuming a matrix W such that $(p, q) = W(x_p, x_e)$, one arrives at the Ricatti matrix equation

$$\dot{W} = A_{21} + A_{22}W - WA_{11} - WA_{12}W$$

with appropriate initial conditions. Here A_{ij} , $i, j = 1, 2$, are the partitions of the matrix in (14) into square quarters. Similarly, since

$$\begin{aligned} p(t) &= W_{11}(t)x_p(t) + W_{12}(t)x_e(t), \\ q(t) &= W_{21}(t)x_p(t) + W_{22}(t)x_e(t), \end{aligned}$$

and since $u = G_q^* p$ and $v = G_e^* q$, the feedback controller is in hand.

In closing, reference is made once more to the analogy between Examples 1 and 2. It is not surprising to find that a distributive system analogy exists for Example 4. While these matters are not dealt with here (see [17]), it suffices to say that (14) remains valid when appropriately interpreted. This equation and the Ricatti equation also have a partial differential form.

4. A sufficient condition for minimax points. Suppose that the functional $f: H \rightarrow R$ introduced in §2 has a second (Fréchet) derivative f'' . Then f'' is a linear operator on H which may be represented in the matrix form

$$[f''(\theta)] = \begin{bmatrix} f''_{11}(\theta) & f''_{12}(\theta) \\ f''_{21}(\theta) & f''_{22}(\theta) \end{bmatrix},$$

where the component second partial derivatives, $f''_{ij}(\theta)$, $i, j = 1, 2$, are linear operators which map according to the table

$$f''_{ij}(\theta): H_i \rightarrow H_j, \quad i, j = 1, 2.$$

This representation is written with the meaning of the equation

$$[f''(\theta)]\psi = (f''_{11}(\theta)\psi_1 + f''_{12}(\theta)\psi_2, f''_{21}(\theta)\psi_1 + f''_{22}(\theta)\psi_2), \quad (\psi_1, \psi_2) \in H.$$

For the specific functional of §3, these second partials are easily located, indeed

$$\begin{aligned}f''_{11}(u, v) &= I + TT^*, \\f''_{12}(u, v) &= -S^*T, \\f''_{21}(u, v) &= -T^*S, \\f''_{22}(u, v) &= SS^* - I.\end{aligned}$$

Consider now the Taylor series expansion of f about the point θ :

$$f(\theta + \delta\theta) = f(\theta) + \langle f'(\theta), \delta\theta \rangle + \langle \delta\theta, [f''(\theta)]\delta\theta \rangle + o(\|\delta\theta\|^2).$$

Suppose that $\theta_0 = (u_0, v_0)$ is a critical point (i.e., $f'(\theta_0) = 0$) and that $\delta\theta = (\delta u, \delta v)$. Then

$$\begin{aligned}f(u_0 + \delta u, v_0) &= f(u_0, v_0) + \langle \delta u, f''_{11}(u_0, v_0)\delta u \rangle, \\f(u_0, v_0 + \delta v) &= f(u_0, v_0) + \langle \delta v, f''_{22}(u_0, v_0)\delta v \rangle,\end{aligned}$$

and the conditions⁴

$$\begin{aligned}f''_{11}(u_0, v_0) &\geq 0, \\f''_{22}(u_0, v_0) &\leq 0\end{aligned}$$

are evidently sufficient to guarantee that (u_0, v_0) is a minimax point. These conditions result in the following lemma.

LEMMA 2. *If the operators $I + TT^*$ and $I - SS^*$ are positive definite, then the critical point (u_0, v_0) is a minimax point for the functional of (3).*

5. Computing the saddle point. Consider once more (4) and (5) which are repeated below in slightly modified form:

$$(16) \quad \lambda = -TT^*\lambda + \xi + SS^*\beta,$$

$$(17) \quad \beta = SS^*\beta + \xi - TT^*\lambda,$$

$$(18) \quad \lambda = (SS^* - TT^*)\lambda + \xi.$$

All three equations are of the form

$$(19) \quad x = Kx - y,$$

where K is a self-adjoint operator. (In the first (second) equation β (respectively, λ) is treated as a parameter.) Thus attention focuses on the invertibility, of an operator $I - K$ on a Hilbert space H .

For the operator K , let $\tilde{N}(K)$ denote the *numerical range* of K , which is, by definition, the set

$$\tilde{N}(K) = \{\langle x, Kx \rangle : \|x\| = 1\}.$$

⁴ The notations $K \geq 0$, $K > 0$, $K \leq 0$ and $K < 0$ denote the property that $\langle x, Kx \rangle \geq 0$, > 0 , ≤ 0 and < 0 , respectively, for all $x \in H$.

Suppose that ϵ is an arbitrary scalar. The distance from ϵ to the set $\tilde{N}(K)$ is defined by

$$d = d[\epsilon; \tilde{N}(K)] = \inf \{ |\epsilon - \mu| : \mu \in \tilde{N}(K) \}.$$

If this distance is strictly greater than zero, then (see Appendix 1) the operator $\epsilon I - K$ is invertible and has a bounded inverse satisfying

$$\|(\epsilon I - K)^{-1}\| \leq \frac{1}{d}.$$

In particular, the case $\epsilon = 1$ provides a sufficient condition for invertibility of the operators in the present discussion.

Because of the particular forms of K , the numerical ranges in each of the three cases is a subset of the real line. In (16), $K = -TT^*$ and

$$\langle x, Kx \rangle = -\langle x, TT^*x \rangle = -\|T^*x\|^2,$$

which implies that $\tilde{N}(-TT^*)$ is a subset of $(-\infty, 0]$. Similarly, one easily shows that $\tilde{N}(SS^*) \subset [0, \infty)$, while the identity

$$\langle x, (SS^* - TT^*)x \rangle = \|S^*x\|^2 - \|T^*x\|^2, \quad x \in H,$$

implies that, in general, $\tilde{N}(SS^* - TT^*)$ may contain both positive and negative points. In the first case, it follows easily that $d[1; \tilde{N}(-TT^*)] \geq 1$ and, consequently, that $I + TT^*$ is invertible with $\|(I + TT^*)^{-1}\| \leq 1$. In the second two cases, the respective numerical ranges may, in general, contain (or have as a limit point) the scalar $\epsilon = 1$. Many different conditions can be imposed to remove this possibility. For instance, if $\|S^*\| < 1$ or if $\mu I - SS^*$ is nonnegative definite for some $\mu < 1$, then in both cases $\epsilon = 1$ is separated from the respective numerical ranges. Consideration of the example $SS^* = s^2I$, $TT^* = t^2I$ for suitable scalars s and t shows, however, that in general $I - SS^*$ may be invertible while $I - SS^* + TT^*$ is not, and conversely.

Assume now that $I - K$ is invertible. Many iterative techniques are available which may be used in the computation of the solution to (18) (see [7], [8] and [9]). In Appendix 1 a rather direct method is suggested. In this method (18) is converted to an equivalent form which involves a contraction operator. The resulting equation can then be solved by direct iteration, the iterates converging at a rate independent of the initial estimate (see [4, p. 27]).

Suppose now that $\{\theta_i : i = 1, 2, \dots\}$ is a sequence which converges to the solution of the equation $f'(\theta_0) = 0$. If the critical point θ_0 is a relative maximum, then each element of the sequence $\{f(\theta_i) : i = 1, 2, \dots\}$ is a lower bound for $f(\theta_0)$. Similarly, for θ_0 a local minima $f(\theta_i)$, $i = 1, 2, \dots$, is an upper bound for $f(\theta_0)$. If, however, the number $f(\theta_0)$ is unknown prior

to the determination of θ_0 , then the differences $f(\theta_i) - f(\theta_0)$, $i = 1, 2, \dots$, cannot be determined and are of little practical use in terminating an iterative sequence.

Rall in [2] has pointed out that minimax points are distinct in that it is possible to generate both upper and lower bounds for the value of f at the critical point. To illustrate in the present setting, rewrite (16) and (17) in the forms

$$(20) \quad \lambda(\beta) = (I + TT^*)^{-1}(\xi + SS^*\beta),$$

$$(21) \quad \beta(\lambda) = (I - SS^*)^{-1}(\xi - TT^*\lambda),$$

respectively. Assume that an estimate β is given and that $\lambda(\beta)$ is the corresponding solution to (20). Similarly for an estimate λ , $\beta(\lambda)$ denotes the corresponding solution to (21). By an abuse of notation, we shall write $f(\lambda, \beta)$ rather than $f(T^*\lambda, S^*\beta)$.

LEMMA 3. *If*

$$\begin{aligned} (SS^*)^{1/2}(I + TT^*)^{-1}(SS^*)^{1/2} &\leq kI \quad \text{for some } k < 1, \\ (TT^*)^{1/2}(I - SS^*)^{-1}(TT^*)^{1/2} &\leq nI \quad \text{for some } n > -1, \end{aligned}$$

then

$$f(\lambda(\beta), \beta) \leq f(\lambda_0, \beta_0) \leq f(\lambda, \beta(\lambda)).$$

Proof. For λ and β independent, a direct expansion of (3) yields

$$\begin{aligned} f(\lambda, \beta) &= \frac{1}{2}\{\|SS^*\beta - TT^*\lambda + \beta\|^2 + \|T^*\lambda\|^2 - \|S^*\beta\|^2\} \\ (22) \quad &= \frac{1}{2}\{\langle\beta, SS^*(SS^* - I)\beta\rangle + \langle\lambda, TT^*(TT^* + I)\lambda\rangle + \langle\xi, \xi\rangle \\ &\quad - 2\langle SS^*\beta, TT^*\lambda\rangle - 2\langle TT^*\lambda, \xi\rangle + 2\langle SS^*\beta, \xi\rangle\}. \end{aligned}$$

For $\beta = \lambda$, this simplifies to

$$\begin{aligned} f(\lambda, \lambda) &= \frac{1}{2}\{\langle\lambda, (TT^* - SS^*)(I + TT^* - SS^*)\lambda\rangle + \langle\xi, \xi\rangle \\ &\quad + 2\langle(SS^* - TT^*)\lambda, \xi\rangle\}, \end{aligned}$$

and when $(I + TT^* - SS^*)\lambda = \xi$, the right-hand side becomes

$$\begin{aligned} \frac{1}{2}\{\langle\xi, \xi\rangle - \langle(TT^* - SS^*)\lambda, \xi\rangle\} &= \frac{1}{2}\{\langle\xi, \xi\rangle - \langle\xi - \lambda, \xi\rangle\} \\ &= \frac{1}{2}\langle\lambda, \xi\rangle. \end{aligned}$$

When $I + TT^* - SS^*$ is invertible, this becomes

$$f(\lambda_0, \beta_0) = \frac{1}{2}\langle(I + TT^* - SS^*)^{-1}\xi, \xi\rangle.$$

Returning now to (22) and letting $(I + TT^*)\lambda = SS^*\beta + \xi$, we have

$$\begin{aligned}
f(\lambda(\beta), \beta) &= \frac{1}{2}\{\langle \beta, SS^*(SS^* - I)\beta \rangle + \langle TT^*\lambda, SS^*\beta + \xi \rangle + \langle \xi, \xi \rangle \\
&\quad - 2\langle SS^*\beta, TT^*\lambda \rangle - 2\langle TT^*\lambda, \xi \rangle + 2\langle SS^*\beta, \xi \rangle\} \\
&= \frac{1}{2}\{\langle \beta, SS^*(SS^* - I)\beta \rangle - \langle TT^*\lambda, SS^*\beta + \xi \rangle + \langle \xi, \xi \rangle \\
&\quad + 2\langle SS^*\beta, \xi \rangle\} \\
&= \frac{1}{2}\{\langle \beta, SS^*(SS^* - I)\beta \rangle + \langle \xi - TT^*\xi, SS^*\beta + \xi \rangle + \langle SS^*\beta, \xi \rangle\} \\
&= \frac{1}{2}\{\langle \lambda, SS^*\beta + \xi \rangle - \langle \beta, SS^*\beta \rangle\},
\end{aligned}$$

and hence, finally,

$$f(\lambda(\beta), \beta) = \frac{1}{2}\{\langle (SS^*\beta + \xi), [I + TT^*]^{-1}(SS^*\beta + \xi) \rangle - \langle \beta, SS^*\beta \rangle\}.$$

Following a similar computation for the case $\beta - SS^*\beta = \xi - TT^*\lambda$ results in the expression

$$f(\lambda, \beta(\lambda)) = \frac{1}{2}\{\langle \beta, \xi - TT^*\lambda \rangle + \langle \lambda, TT^*\lambda \rangle\},$$

and when $I - SS^*$ is invertible,

$$f(\lambda, \beta(\lambda)) = \frac{1}{2}\{\langle (\xi - TT^*\lambda), [I - SS^*]^{-1}(\xi - TT^*\lambda) \rangle + \langle \lambda, TT^*\lambda \rangle\}.$$

To complete the proof it remains only to compare the above three quantities. In Lemma 1 of Appendix 2 and its corollary it is shown that

$$\begin{aligned}
f(\lambda_0, \beta_0) - f(\lambda(\beta), \beta) &= \langle \omega(\beta), [(I + TT^* - SS^*)^{-1} - (I + TT^*)^{-1}]\omega(\beta) \rangle, \\
f(\lambda, \beta(\lambda)) - f(\lambda_0, \beta_0) &= \langle \omega(\lambda), [(I + TT^* - SS^*)^{-1} - (I - SS^*)^{-1}]\omega(\lambda) \rangle,
\end{aligned}$$

where ω is the function $\omega(x) = -\xi + (I + TT^* - SS^*)x$. Moreover, Lemma 2 of Appendix 2 and its corollary show that the hypothesis of the present lemma is sufficient to assure that

$$(I + TT^* - SS^*)^{-1} - (I + TT^*)^{-1} \geq 0,$$

while

$$(I + TT^* - SS^*)^{-1} - (I - SS^*)^{-1} \leq 0.$$

Remark 5. If $\beta = \beta_0 + \delta\beta$, where $\beta_0 = (I + TT^* - SS^*)^{-1}\xi$, then

$$\omega(\beta) = -\xi + (I + TT^* - SS^*)^{-1}\beta = (I + TT^* - SS^*)^{-1}\delta\beta.$$

Using the results of Appendix 2, we then have that

$$\begin{aligned}
f(\lambda_0, \beta_0) - f(\lambda(\beta), \beta) &= \langle \delta\beta, Q\delta\beta \rangle, \\
f(\lambda, \beta(\lambda)) - f(\lambda_0, \beta_0) &= -\langle \delta\beta, R\delta\beta \rangle,
\end{aligned}$$

where $Q \geq 0$ and $R \geq 0$.

Remark 6. It is interesting to note that the condition $\|S\| < 1$ is sufficient to guarantee that the hypothesis of the above lemma is satisfied. Indeed, note that $(I + TT^*)^{-1} < I$ and hence that (recall $\|S\| = \|S^*\|$)

$$\langle x, (SS^*)^{1/2}(I + TT^*)^{-1}(SS^*)^{1/2}x \rangle = \langle (SS^*)^{1/2}x, (I + TT^*)^{-1}(SS^*)^{1/2}x \rangle \\ \leq \| (SS^*)^{1/2}x \|^2 = \langle x, SS^*x \rangle = \| S^*x \|^2 \leq \| x \|^2.$$

In the second case, $\|S\| < 1$ implies $I - SS^* > 0$ and the hypothesis follows trivially. By the same token, the condition $\|S\| < 1$ also guarantees the invertibility of $I + TT^* - SS^*$, and the sufficiency condition of Lemma 2 in §4.

6. In closing. In §§3, 4 and 5 attention is centered on the minimax points of a particular functional. The principal results are embodied in Lemma 1 and its corollary, Lemma 2 and Lemma 3. The four examples of §3 illustrate the corollary of Lemma 1 in a selection of concrete settings. Sections 4 and 5 deal with sufficiency conditions and some computational aspects of the minimax problem.

Remark 1 indicates some generalizations which extend the scope of the results. It is interesting that the results are also useful in some types of constrained minimax problems. To illustrate, consider two types of constraints on the control $u \in H_1$, namely,

$$\begin{aligned} \zeta - Au &= 0, \\ \|u\|^2 &\leq K^2. \end{aligned}$$

Here A is a linear operator with values in a Hilbert space wherein ζ is a fixed element. The first (linear) constraint might typically arise in specifying terminal conditions on the admissible controls. The second constraint is obviously a maximum pursuer energy limitation.

To determine the constrained critical points for the functional of (3), the well-known Lagrange multiplier technique (see [4, §47]) is employed with the resulting functionals:

$$\begin{aligned} \frac{1}{2}\{\|Sv - Tu + \xi\|^2 + \|u\|^2 - \|v\|^2 + 2\langle \alpha, \zeta - Au \rangle\}, \\ \frac{1}{2}\{\|Sv - Tu + \xi\|^2 + \|u\|^2 - \|v\|^2 + \gamma(\|u\|^2 + p^2 - k^2)\}, \end{aligned}$$

in the two cases respectively. Here γ and α are appropriate Lagrange multipliers, while p is an artificial variable introduced to cope with the inequality in the second constraint.

The first functional may be rewritten as

$$\frac{1}{2}\{\|Sv - Tu + \xi\|^2 + \|u - A^*\zeta\|^2 - \|v\|^2\} + \frac{1}{2}\{\langle \alpha, \zeta \rangle - \|A^*\zeta\|^2\}.$$

The first part is of a form considered in Remark 1 and the second part is

independent of u and v . Similarly, the second functional can be regrouped into a term considered in Remark 1 plus a term independent of u and v . Consequently, the results of §3 may be used directly. Naturally the situation is complicated by the appearance of parameters in the solution equations which must be determined from the added constraints.

Appendix 1.

THEOREM. *Let K be everywhere defined on H and suppose that the scalar 1 is at a positive distance d from $\tilde{N}(K)$. Then*

$$(A1.1) \quad x = Kx - \xi$$

has a unique solution for every $\xi \in H$; the operator $K - I$ is invertible and $\|(K - I)^{-1}\| \leq d$. Moreover, for $0 < \alpha < 2d/\|K - I\|^2$ and arbitrary $x_0 \in H$,

$$(K - I)^{-1}\xi = \lim_{n \rightarrow \infty} [(1 - \alpha)I + \alpha(K - \xi)]^n x_0,$$

where $K - \xi$ is the operator mapping x into $Kx - \xi$.

Proof. The first part of the theorem is a classic (see [18, p. 147] or [19, p. 1]) and will not be proved here. The second part is not quite so usual and the proof of Zarantonello [19] is included here for completeness. First note that, for $\alpha > 0$,

$$x = [\alpha K + (1 - \alpha)I]x - \alpha\xi$$

is equivalent to $x = Kx - \xi$, and hence it suffices to show that $\alpha K + (1 - \alpha)I$ is a contraction. Consider the expansion

$$\|[\alpha K + (1 - \alpha)I]x\|^2 = \|x\|^2 - 2\alpha \langle (K - I)x, x \rangle + \alpha^2 \|(K - I)x\|^2.$$

Using the definition of d we obtain

$$\|[\alpha K + (1 - \alpha)I]x\|^2 \leq (1 - 2\alpha d + \alpha^2 \|K - I\|^2) \|x\|^2,$$

from which it appears that if $\alpha < 2d/\|K - I\|^2$, then

$$\|\alpha K + (1 - \alpha)I\| < 1,$$

which completes the proof.

If α is chosen to minimize the right-hand side of (A1.1), the result is the obvious corollary.

COROLLARY 1. *For $\alpha = d/\|K - I\|$,*

$$\|\alpha K + (1 - \alpha)I\|^2 \leq 1 - \left(\frac{d}{\|K - I\|} \right)^2.$$

Appendix 2.

LEMMA 1. *Whenever the indicated inverses exist, the identity*

$$\begin{aligned} \langle \xi, (I + TT^* - SS^*)^{-1}\xi \rangle + \langle \beta, SS^*\beta \rangle \\ - \langle (SS^*\beta + \xi), (I + TT^*)^{-1}(SS^*\beta + \xi) \rangle \\ = \langle \omega, [(I + TT^* - SS^*)^{-1} - (I + TT^*)^{-1}]\omega \rangle, \end{aligned}$$

where $\omega = -\xi + (I + TT^* - SS^*)\beta$, holds for all β and ξ .

Proof. Consider first the identity

$$\begin{aligned} \langle (SS^*\beta + \xi), (I + TT^*)^{-1}(SS^*\beta + \xi) \rangle \\ = \langle \omega, (I + TT^*)^{-1}\omega \rangle - \langle \beta, (I + TT^*)\beta \rangle - 2\langle \beta, \xi + SS^*\beta \rangle, \end{aligned}$$

which may be verified by expansion of the right-hand side. Using this expansion the left-hand side of the identity of the lemma becomes

$$\begin{aligned} \langle \xi, (I + TT^* - SS^*)^{-1}\xi \rangle + \langle \beta, SS^*\beta \rangle - \langle \omega, (I + TT^*)^{-1}\omega \rangle \\ + \langle \beta, (I + TT^*)\beta \rangle - 2\langle \beta, \xi + SS^*\beta \rangle, \end{aligned}$$

which reduces to

$$\begin{aligned} \langle \xi, (I + TT^* - SS^*)^{-1}\xi \rangle + \langle \beta, (I + TT^* - SS^*)\beta \rangle - 2\langle \beta, \xi \rangle \\ - \langle \omega, (I + TT^*)^{-1}\omega \rangle. \end{aligned}$$

The lemma now follows directly from the identity

$$\begin{aligned} \langle \xi, (I + TT^* - SS^*)^{-1}\xi \rangle + \langle \beta, (I + TT^* - SS^*)\beta \rangle - 2\langle \beta, \xi \rangle \\ = \| (I + TT^* - SS^*)^{1/2}\beta - (I + TT^* - SS^*)^{-1/2}\xi \|^2 \\ = \| (I + TT^* - SS^*)^{-1/2}\omega \|^2 \\ = \langle \omega, (I + TT^* - SS^*)^{-1}\omega \rangle. \end{aligned}$$

COROLLARY. *Whenever the indicated inverses exist,*

$$\begin{aligned} \langle \xi, (I + TT^* - SS^*)^{-1}\xi \rangle - \langle \lambda, TT^*\lambda \rangle \\ - \langle (TT^*\lambda - \xi), (I - SS^*)^{-1}(TT^*\lambda - \xi) \rangle \\ = \langle \omega, [(I + TT^* - SS^*)^{-1} - (I - SS^*)^{-1}]\omega \rangle, \end{aligned}$$

where $\omega = -\xi + (I + TT^* - SS^*)\lambda$.

Proof. The corollary follows directly from the lemma as a result of the changes of variables: $SS^* \rightarrow -TT^*$, $TT^* \rightarrow SS^*$, $\beta \rightarrow \lambda$.

LEMMA 2. *If $I + TT^* - SS^*$ and $I - SS^*$ are invertible, then*

$$\begin{aligned} (I + TT^* - SS^*)^{-1} - (I + TT^*)^{-1} \\ \text{(A2.1)} \quad = (I + TT^*)^{-1}(SS^*)^{1/2}[I - (SS^*)^{1/2}(I + TT^*)^{-1}(SS^*)^{-1/2}]^{-1} \\ \cdot (SS^*)^{1/2}(I + TT^*)^{-1}, \end{aligned}$$

$$\begin{aligned}
 & (I + TT^* - SS^*)^{-1} - (I - SS^*)^{-1} \\
 (A2.2) \quad & = (I - SS^*)^{-1}(TT^*)^{1/2}[I + (TT^*)^{1/2}(I - SS^*)^{-1}(TT^*)^{1/2}]^{-1} \\
 & \quad \cdot (TT^*)^{1/2}(I - SS^*)^{-1}.
 \end{aligned}$$

Proof. These equalities may be established by clearing fractions. To illustrate, set $M = I + TT^*$ and $N = SS^*$ in which case (A2.1) becomes

$$\begin{aligned}
 (M - N)^{-1} - M^{-1} &= M^{-1}N^{1/2}[I - N^{1/2}M^{-1}N^{1/2}]^{-1}N^{1/2}M^{-1} \\
 &= M^{-1}N(I - M^{-1}N)^{-1}M^{-1}.
 \end{aligned}$$

Since $(M - N)^{-1} = [M(I - M^{-1}N)]^{-1} = (I - M^{-1}N)^{-1}M^{-1}$, multiplication by M on the right produces

$$(I - M^{-1}N)^{-1} - I = M^{-1}N(I - M^{-1}N)^{-1}.$$

Clearing fractions then completes the proof.

COROLLARY. *If $(SS^*)^{1/2}(I + TT^*)^{-1}(SS^*)^{1/2} \leq xI$ for some $x < 1$, then*

$$(A2.3) \quad (I + TT^* - SS^*)^{-1} - (I + TT^*)^{-1} \geq 0.$$

If $(TT^)^{1/2}(I - SS^*)^{-1}(TT^*)^{1/2} \geq xI$ for some $x > -1$, then*

$$(A2.4) \quad (I + TT^* - SS^*)^{-1} - (I - SS^*)^{-1} \leq 0.$$

Proof. If A is any linear operator and if $K \geq 0$, then $A^*KA \geq 0$. Thus, in proving (A2.3), it suffices to consider $[I - (SS^*)^{1/2}(I + TT^*)^{-1}(SS^*)^{1/2}]^{-1}$. The hypothesis of the corollary guarantees that $I - (SS^*)^{1/2}(I + TT^*)^{-1} \cdot (SS^*)^{1/2}$ is positive definite and has a positive definite inverse. The proof of (A2.4) is analogous.

REFERENCES

- [1] Y. C. HO, A. BRYSON AND S. BARON, *Differential games and optimal pursuit-evasion strategies*, IEEE Trans. Automatic Control, AC-10 (1965), pp. 385-389.
- [2] L. B. RALL, *On complementary variational principles*, J. Math. Anal. Appl., 14 (1966), pp. 174-184.
- [3] W. A. PORTER, *Modern Foundations of System Engineering*, Macmillan, New York, 1966.
- [4] L. LIUSTERNIK AND V. SOBOLEV, *Elements of Functional Analysis*, Ungar, New York, 1961.
- [5] M. M. VAINBERG, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, San Francisco, 1964.
- [6] G. F. SIMMONS, *Topology and Modern Analysis*, McGraw-Hill, New York, 1963.
- [7] M. ALTMAN, *Approximation Methods in Functional Analysis*, Lecture notes, California Institute of Technology, Pasadena, 1959.
- [8] P. ANSELONE, ed., *Nonlinear Integral Equations*, University of Wisconsin Press, Madison, 1964.
- [9] L. KANTOROVICH AND G. AKILOV, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.

- [10] H. HSIEH AND C. LEONDES, *Modern Control Systems Theory*, McGraw-Hill, New York, 1965, Chap. II.
- [11] I. MCCausLAND, *On optimum control of temperature distribution in a solid*, J. Electronics Control, 14 (1963), p. 655.
- [12] P. K. C. WANG, *Control of distributive parameter systems*, Advances in Control Systems, vol. 1, C. T. Leondes, ed., Academic Press, New York, 1964, pp. 75-171.
- [13] A. BALAKRISHNAN, *On the state space theory of linear systems*, J. Math. Anal. Appl., 14 (1966), pp. 371-391.
- [14] R. GOLDBERG, *Fourier Transforms*, Cambridge University Press, London, 1962.
- [15] S. BOCHNER AND K. CHANDRASEKHARAN, *Fourier Transforms*, Princeton University Press, Princeton, 1949.
- [16] H. TSIEH, *Engineering Cybernetics*, McGraw-Hill, New York, 1954.
- [17] M. FAHMY, *A solution technique for a class of optimal control problems in distributive systems*, Doctoral dissertation, University of Michigan, Ann Arbor, 1965.
- [18] M. STONE, *Linear Transformations in Hilbert Space*, Colloquium Publications, vol. XV, American Mathematical Society, New York, 1932.
- [19] E. ZARANTONELLO, *The closure of the numerical range contains the spectrum*, Tech. Rep. 7, Department of Mathematics, University of Kansas, Lawrence, 1964.

ON THE CONTROLLABILITY OF DELAY-DIFFERENTIAL SYSTEMS*

LEONARD WEISS†

1. Introduction. The importance of dealing effectively with the inevitable delays of signal transmission within a control system is attested to by the volume of literature devoted to this problem over the years [1]. The early textbooks on control generally treated the problem of time lags by ad hoc and approximation methods, some of which involved modeling a system with pure delay by a higher order system without pure delay (see Repin [2] for a detailed discussion of this technique).

For a wide class of systems, however, it is natural and important that the model show the delay explicitly (see [3], [4]), which motivates the consideration of delay-differential equations as models and the study of their properties from a system-theoretic point of view.

One of the fundamental system-theoretic properties of a control system is that of "controllability", which can be viewed as pertaining to the question of whether a given (optimal) control problem is well posed or not, and which therefore impinges on questions of existence of solutions to such problems. Exactly how one should define the concept of controllability depends on the nature of the problem one is considering. Even in the case of control systems with finite-dimensional state spaces, there is more than one natural way of defining controllability [5]. In the case of infinite-dimensional spaces and with possibly infinite-dimensional target sets, the controllability concept of interest certainly depends on the precise nature of the target set.

In this paper we define and discuss a type of controllability which is likely to play an important role in a broad class of optimal control problems for systems described by delay-differential equations. One of our objectives is to illustrate that some techniques which have been found to be eminently useful in obtaining results for ordinary differential equations can also be profitably used when dealing with delay equations. In particular, the approach we take to the solution of the problem discussed in the sequel is analogous to that for ordinary differential equations given by Markus and Lee [6] as modified by Kalman [7]. The results subsume the controllability results given by Chyung and Lee [8] in their paper on opti-

* Received by the editors April 11, 1967, and in revised form July 27, 1967.

† Center for Dynamical Systems, Brown University, Providence, Rhode Island. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, United States Air Force, under AFOSR Grants 68-1346 and 693-66, in part by the Alfred P. Sloan Foundation in the form of a Research Fellowship to the author, and in part by the National Aeronautics and Space Administration under Grant NGR-40-002-015.

mal control of delay-differential systems with target sets in Euclidean space.

2. Definition of controllability and some preliminary remarks. Consider the system

$$(1) \quad \dot{x}(t) = f(t, x(t), x(t-h), u(t)), \quad t > t_0,$$

where $x(t) \in R^n$, $u(t) \in R^p$ and u is measurable and bounded on every finite time interval,¹ h is a positive constant (the delay), $f \in C^1$ in all its arguments and $f(t, 0, 0, 0) \equiv 0$. Let \mathfrak{B} be the Banach space of real n -vector-valued continuous functions defined on the interval $[t_0 - h, t_0]$ with the uniform norm, i.e., if $\phi \in \mathfrak{B}$, we have $\|\phi\| = \max_{t \in [t_0 - h, t_0]} |\phi(t)|$. Then a solution of (1) exists and is unique for $t > t_0$ if one specifies an initial function $\phi \in \mathfrak{B}$ (see [9]).

Remark. The assumption of a single constant delay is for convenience only. All the results in this paper can be easily generalized to the case of multiple delays, and these delays can also be time-varying as long as they are appropriately bounded so that their values do not overlap.

Let \mathfrak{C} be an abstract normed linear space of functions defined on the interval $[t_0 - h, t_0]$. Then we give the following definitions.

DEFINITION 1. A system (1) is *controllable to a function* $\psi(\cdot) \in \mathfrak{C}$ with respect to the space of initial functions \mathfrak{B} if, for any given $\phi \in \mathfrak{B}$, there exist a time t_1 , $t_0 < t_1 < \infty$, and an admissible control segment² $u_{[t_0, t_1+h]}$ such that $x(t; t_0, \phi, u) = \psi(t - t_1 + t_0 - h)$, $t \in [t_1, t_1 + h]$, where $x(t; t_0, \phi, u)$ is the solution of (1), starting at time t_0 , with initial function ϕ and control u .

DEFINITION 2. If the system (1) is controllable to all functions in \mathfrak{C} , it is *controllable to the space* \mathfrak{C} .

DEFINITION 3. If $\psi(\cdot) \equiv 0$ in Definition 1, then the system is *controllable to the origin*.

DEFINITION 4. If t_1 is constant with ϕ in any of the above definitions, the corresponding type of controllability is *uniform*.

In the sequel, we shall give sufficient conditions for (1) to be *controllable to the origin as well as to a function with respect to the space* \mathfrak{B} . We shall also give sufficient conditions under which the linear system

$$(2) \quad \dot{x}(t) = A(t)x(t) + B(t)x(t-h) + C(t)u(t),$$

where $x(t) \in R^n$, $u(t) \in R^p$, and $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ are continuous matrix functions, is controllable to the origin and to a function with respect to \mathfrak{B} . The aforementioned conditions for (2) will be shown to be necessary if a

¹ Such functions will be referred to as "admissible".

² A segment $g_{[a, b]}$ denotes a function g defined over the interval $[a, b]$.

certain assumption about the space of trajectories of the homogeneous equation

$$(3) \quad \dot{x}(t) = A(t)x(t) + B(t)x(t-h)$$

is true.

It should be strongly emphasized that controllability to the origin for a delay-differential system does *not* imply, in general, controllability to a function or a space of functions. However, the techniques which are used in this paper to study controllability to the origin are completely applicable to the study of controllability to a function or function space. This fact is illustrated in §6, where some results along this line are given.

3. The linear problem. Consider (2) with \mathfrak{B} the space of initial functions. The solution of (2) for time $t > t_0$, and corresponding to initial function $\phi \in \mathfrak{B}$, has the form (see [10]):

$$(4) \quad x(t) \equiv x(t; t_0, \phi, u) = M(t, t_0, \phi) + \int_{t_0}^t K(s, t)C(s)u(s) ds,$$

where $M(t, t_0, \phi)$ is the solution of the homogeneous equation (3) corresponding to initial time t_0 and initial function ϕ , i.e.,

$$(5) \quad M(t, t_0, \phi) = \phi(t) \quad \text{for } t \in [t_0 - h, t_0].$$

The kernel $K(s, t)$ is defined for $t \geq t_0$ and $t_0 \leq s \leq t$ and is an $n \times n$ matrix solution of the equations:

$$(6a) \quad \frac{\partial K(s, t)}{\partial s} = -K(s, t)A(s) - K(s+h, t)B(s+h),$$

$$t_0 \leq s \leq t-h,$$

$$(6b) \quad \frac{\partial K(s, t)}{\partial s} = -K(s, t)A(s),$$

$$t-h \leq s \leq t,$$

with $K(t, t) = I$ (the identity matrix).

Equation (6b) shows the obvious fact that over one delay interval, the delay equation behaves similarly to an ordinary differential equation with $K(s, t)$ playing the role of a fundamental matrix solution of the homogeneous equation (see [11]).

LEMMA 1. *Let (2) be given with any $\phi \in \mathfrak{B}$. A sufficient condition for existence of an admissible control which results in the solution having a zero-crossing in finite time is that there exists $t_1 > t_0$ such that*

$$(7) \quad \text{rank} \int_{t_0}^{t_1} K(s, t_1)C(s)C'(s)K'(s, t_1) ds = n,$$

where ' indicates transpose.

Proof. Let $\mathfrak{C}(t_0, t_1) = \int_{t_0}^{t_1} K(s, t_1)C(s)C'(s)K'(s, t_1) ds$. In (4), substitute

$$u(s) = -C'(s)K'(s, t_1)\mathfrak{C}(t_0, t_1)^{-1}M(t_1, t_0, \phi).$$

Then $x(t_1) = 0$.

DEFINITION 5. *The force-free attainable set at time t of a system (2) is the set of all points in R^n that can be reached at time t by the trajectories of (3) resulting from all initial functions contained in \mathfrak{B} .*

DEFINITION 6. A system (2) whose force-free attainable set at any time t is all of R^n is *pointwise complete*.

Since we have been unable to give an example to the contrary, we present for the reader's amusement the following conjecture.

CONJECTURE. All constant coefficient systems of the form (2) are pointwise complete.

Remark. The conjecture is true if we consider the trajectories only on the interval $t_0 - h \leq t \leq t_0 + h$, since the elements of \mathfrak{B} span all of R^n at any time $t \in [t_0 - h, t_0]$ and the system (3) behaves as an ordinary differential equation on the interval $[t_0, t_0 + h]$.

LEMMA 2. *If a system (2) is pointwise complete, then (7) is necessary as well as sufficient for existence of a control which results in a zero-crossing in finite time of the solution of (2) for any $\phi \in \mathfrak{B}$.*

Proof. Given any $\phi \in \mathfrak{B}$, suppose there exist $t_1 > t_0$ and a control $u_{[t_0, t_1]}$ such that $x(t_1) = 0$, but (7) does not hold. The latter implies that there exists a nonzero vector $x_1 \in R^n$ such that $x_1'K(s, t_1)C(s) = 0, t_0 \leq s \leq t_1$. Then, from (4), $x_1'M(t_1, t_0, \phi) = 0$. By hypothesis, however, ϕ can be chosen so that $M(t_1, t_0, \phi) = x_1$. Then $x_1'x_1 = 0$ which contradicts the assumption that $x_1 \neq 0$.

THEOREM 1. *A pointwise complete system (2) is controllable to the origin with respect to \mathfrak{B} if and only if*

(i) *there exists $t_1 > t_0$ such that (7) holds;*

(ii) *given $\phi \in \mathfrak{B}$, then with t_1 as in (7) and for some admissible $u_{[t_0, t_1]}$ such that $x(t_1; t_0, \phi, u_{[t_0, t_1]}) = 0$, the equation*

$$(8) \quad C(t)u(t) = -B(t)x(t - h; t_0, \phi, u_{[t_0, t_1]})$$

has an admissible solution $u(\cdot)$ on the interval $(t_1, t_1 + h)$.

Proof. By Lemma 1 we have that for any $\phi \in \mathfrak{B}$ there exists $u_{[t_0, t_1]}$ such that $x(t_1; t_0, \phi, u_{[t_0, t_1]}) = 0$. If (8) holds, then over the interval $(t_1, t_1 + h)$, (2) becomes

$$(9) \quad \dot{x}(t) = A(t)x(t), \quad x(t_1) = 0.$$

It follows by the uniqueness theorem for ordinary differential equations that $x(t) = 0$ for all $t \in [t_1, t_1 + h]$.

Conversely, if (2) is controllable to the origin with respect to \mathfrak{B} , then for any $\phi \in \mathfrak{B}$, there exist $t_1 > t_0$ and an admissible control $u_{[t_0, t_1+h]}$ such that $x(t; t_0, \phi, u_{[t_0, t_1+h]}) = 0$ for all $t \in [t_1, t_1 + h]$ which implies (8). Since $x(t_1; t_0, \phi, u_{[t_0, t_1]}) = 0$ and the system is pointwise complete, (7) must hold by Lemma 2.

Remark. If the control $u_{[t_0, t_1+h]}$ transfers an initial function $\phi \in \mathfrak{B}$ of the system (2) to the origin (the zero function on the interval $[t_1, t_1 + h]$), then if $u(t) = 0$ for all $t > t_1 + h$, the system will remain at the origin.

4. On the solution of (8). Consider the following facts:

I. An admissible solution of (8) will exist on an interval $(t_1, t_1 + h)$ if and only if $-B(t)x(t - h; t_0, \phi, u_{[t_0, t_1]})$ is in the range of $C(t)$ almost everywhere on $(t_1, t_1 + h)$. Standard techniques can then be employed to construct a solution (see [12]).

II. If "controllable" is replaced by "uniformly controllable" in Theorem 1, then the right side of (8) must be in the range of $C(t)$ for all $\phi \in \mathfrak{B}$ on $(t_1, t_1 + h)$, where t_1 is fixed.

III. No solution of (8) can be unique since one can add to it any vector-valued function of time which is in the null-space of $C(\cdot)$ almost everywhere on $(t_1, t_1 + h)$.

To obtain sharper results than the preceding, it is necessary to do some deep analysis of the attainable set for (2), as indicated by the results below.

Consider (8) over an interval $(t_1, t_1 + h)$ and let P be the set of initial functions in \mathfrak{B} which are controllable to the origin using admissible controls defined over $[t_0, t_1 + h]$. ($P = \mathfrak{B}$ for uniform controllability.) For each $\phi \in P$, let $K_\phi = \{u_\lambda^\phi, \lambda \in \Lambda(\phi)\}$ = the set of admissible controls taking ϕ to the origin (the zero function defined over the fixed time interval $[t_1, t_1 + h]$). Invoking the axiom of choice, define

$$Q = \{\psi; \psi: P \rightarrow \bigcup_{\phi \in P} K_\phi\}$$

(i.e., $\psi \in Q$ implies $\psi(\phi) = u_\lambda^\phi$ for some $\lambda \in \Lambda(\phi)$). Now, let

$$S_\psi(t) = \{x(t; t_0, \phi, \psi(\phi)); \phi \in P\},$$

where $x(t; t_0, \phi, \psi(\phi))$ denotes the value at time t of the trajectory of (2) generated by initial function ϕ and control $\psi(\phi)$. We then have the following lemma.

LEMMA 3. *If for each $\psi \in Q$ and each $t \in (t_1 - h, t_1)$ the set $S_\psi(t)$ covers all directions in Euclidean n -space, then a necessary and sufficient condition for (8) to have a solution regardless of $u_{[t_0, t_1]}$ almost everywhere on $(t_1, t_1 + h)$ is that there exists a $p \times n$ matrix $D(t)$ with bounded measurable elements such that $B(t) = C(t)D(t)$ almost everywhere on $(t_1, t_1 + h)$.*

Proof. Fix $t \in (t_1, t_1 + h)$. The problem reduces to solving the algebraic

equation

$$Cu = -Bx,$$

where x is an n -vector which can take on values corresponding (except for a magnitude constraint) to any collection of n basis vectors. Then $-Bx \in \text{range } C$ if and only if the columns of B are linear combinations of those of C , i.e., there exists D such that $B = CD$. Continuity of $B(t)$ and $C(t)$ assure that this process can be repeated for each $t \in (t_1, t_1 + h)$ with the matrix $D(t)$ having bounded measurable elements on that interval.

Remark. Under the above conditions, the solution for $u(\cdot)$ has the form

$$(10) \quad u(t) = \sum \alpha_i e_i(t) + D(t)x(t - h; t_0, \phi, u_{[t_0, t_1]}), \quad t_1 < t < t_1 + h,$$

where $e_i(t) \in \text{null-space of } C(t)$ and $\alpha_i = \text{const.}$ The preceding facts together with Theorem 1 immediately imply that the following theorem holds.

THEOREM 2. *A pointwise complete system (2), which satisfies the hypothesis of Lemma 3 is uniformly controllable to the origin with respect to \mathfrak{B} if and only if*

(i) *there exists $t_1 > t_0$ such that (7) holds;*

(ii) *there exists an $n \times p$ matrix, $D(t)$, with bounded measurable elements such that, with t_1 defined as above, $B(t) = C(t)D(t)$ almost everywhere on $(t_1, t_1 + h)$.*

Since engineers have an aversion (and rightfully so!) to measurable solutions of control problems, we give the conditions under which one can find an absolutely continuous solution to (8) over the interval $(t_1, t_1 + h)$. The result emerges as an application of the next lemma which is due to Doležal [13].³

LEMMA 4 (Doležal). *Let $G(t)$ be an $n \times p$ matrix defined on an interval $[a, b]$ and continuous, at least. Suppose there exists an integer $r \leq p$ such that $\text{rank } G(t) = r$ for all $t \in [a, b]$. Then there exists a $p \times p$ matrix $H(t)$, defined and nonsingular on $[a, b]$ and whose degree of smoothness matches that of $G(t)$, such that*

$$G(t)H(t) = [F(t):0], \quad t \in [a, b],$$

where $F(t)$ is $n \times r$, $\text{rank } F(t) = r$ for all $t \in [a, b]$.

THEOREM 3. *If (8) has an admissible solution and if $\text{rank } C(t) = r = \text{const.}$ for all $t \in [t_1, t_1 + h]$, then that solution can be chosen to be absolutely continuous.*

Proof. By Lemma 4, there exist real n -vector-valued continuous functions $c_1(t), \dots, c_r(t)$ which span the range $C(t)$ at each $t \in [t_1, t_1 + h]$. Then,

³ This important lemma has a variety of applications to problems in system theory [16], [17].

if (8) has a solution almost everywhere on $(t_1, t_1 + h)$, we can write

$$(11) \quad B(t)x(t - h; t_0, \phi, u_{[t_0, t_1]}) = \sum_{i=1}^r \alpha_i(t)C_i(t) \quad \text{a.e. on } (t_1, t_1 + h).$$

But since the left side of (11) is absolutely continuous, the α_i 's can be chosen to be absolutely continuous. It then follows that an absolutely continuous solution of (8) exists.

5. The nonlinear problem. The problem will be solved in two steps. First, conditions are given under which one can control a system (1) to an arbitrarily small neighborhood of the origin in finite time, and then we give conditions under which the origin can be reached in finite time from a point in its neighborhood.

DEFINITION 6. A system (1) is *quasi-controllable to the origin with respect to* \mathfrak{B} if for any $\phi \in \mathfrak{B}$ and any $\epsilon > 0$, there exist $t_1 > t_0$ and an admissible control $u_{[t_0, t_1+h]}$ such that

$$\|x(\cdot; t_0, \phi, u)\|_{[t_1, t_1+h]} = \max_{t_1 \leq t \leq t_1+h} |x(t; t_0, \phi, u)| < \epsilon.$$

Consider the system (1) with $f(t, 0, 0, 0) \equiv 0$, $f \in C^1$ in $R \times R^n \times R^n \times R^p$, $u(t) \in R^p$, and $\phi \in \mathfrak{B}$.

Define the following functions:

$\omega(\cdot)$ is a continuous, real-valued nondecreasing function such that

$\omega(s) > s$, $s > 0$;

$\mu(\cdot)$ and $\nu(\cdot)$ are continuous, real-valued functions of s defined for $s \geq 0$, and positive and nondecreasing for $s \neq 0$;

$\beta(\cdot)$ is a continuous, real-valued function of s defined for $s \geq 0$, and positive for $s \neq 0$.

THEOREM 4. *Given are the system (1) and the above defined quantities. Suppose there exist a real-valued function $V(t, x)$, defined and continuous for $t \geq t_0 - h$, $x \in R^n$, and a real p -vector-valued function $U(x)$ which is C^1 in R^n such that*

$$(i) \quad \mu(|x|) \leq V(t, x) \leq \nu(|x|), \quad t \geq t_0 - h,$$

$$(ii) \quad \frac{\partial V(t, x)}{\partial t} \Big|_{x=\rho(t)} + \frac{\partial V(t, x)}{\partial x} \Big|_{x=\rho(t)} \cdot f(t, \rho(t), \rho(t-h), U(\rho(t))) \\ \leq -\beta(|\rho(t)|)$$

for all $t \geq t_0$ and all continuous, real n -vector-valued function segments $\rho_{[t-h, t]}$ such that

$$(iii) \quad V(\xi, \rho(\xi)) < \omega(V(t, \rho(t))), \quad t - h \leq \xi \leq t.$$

Then the system (1) is quasi-controllable with respect to \mathfrak{B} .

Remark. Theorem 4 is an easy generalization of a theorem originally due to Krasovskii [14] on uniform asymptotic stability of delay-differential equations. The proof follows precisely the novel but lengthy proof given by

Driver [9] of the original theorem and will therefore not be reproduced here. Let it suffice to say that if the conditions of the theorem are met, then for any initial-function $\phi \in \mathfrak{B}$, there exists an admissible control which has the effect of driving the system to an ϵ -neighborhood of the origin (in function space) in finite time.

Now, consider the following definitions.

DEFINITION 7. A system (1) is *locally controllable to the origin with respect to \mathfrak{B}* if it is controllable to the origin with respect to a neighborhood $N(0^{\mathfrak{B}})$ of the origin in \mathfrak{B} .

DEFINITION 8. The *first variation of (1) about the zero-solution* is the system (2) where

$$\begin{aligned} A(t) &= \frac{\partial f}{\partial x}(t, 0, 0, 0), \\ B(t) &= \frac{\partial f}{\partial x_d}(t, 0, 0, 0), \quad x_d(t) = x(t - h), \\ C(t) &= \frac{\partial f}{\partial u}(t, 0, 0, 0). \end{aligned}$$

THEOREM 5. A system (1) is locally controllable to the origin with respect to \mathfrak{B} if its first variation about the zero-solution satisfies the conditions:

- (i) there exists $t_1 > t_0$ such that (7) holds;
- (ii) with t_1 defined as above, there exists an $n \times p$ matrix $D(t)$ with bounded, measurable elements such that $B(t) = C(t)D(t)$ almost everywhere on $(t_1, t_1 + h)$.

Proof. We introduce a parameter ξ into the control u and define

$$(12) \quad u^{\xi}(t) = u(t, \xi) = \begin{cases} C'(t)K'(t, t_1)\xi & \text{for } t_0 \leq t \leq t_1, \\ \text{solution}^4 \text{ of } C(t)u(t) = -B(t)x(t - h); \\ & t_0, 0^{\mathfrak{B}}, u^{\xi}) & \text{for } t_1 < t < t_1 + h. \end{cases}$$

Note 1. $u(t, 0) = u^0(t) = 0$ for $t \in [t_0, t_1]$.

Note 2. If $\phi \equiv 0$, then $x(t; t_0, 0^{\mathfrak{B}}, u^0) = 0$ on $[t_0 - h, t_1]$.

Let

$$(13) \quad J(t) = \left. \frac{\partial x(t; t_0, 0^{\mathfrak{B}}, u^{\xi})}{\partial \xi} \right|_{\xi=0}.$$

Since $\phi \equiv 0$, the solution of (1) is written as

$$x(t; t_0, 0^{\mathfrak{B}}, u^{\xi}) \equiv x(t; \xi) = \int_{t_0}^t f(\tau, x(\tau), x_d(\tau), u^{\xi}(\tau)) d\tau, \\ t_0 \leq t \leq t_1 + h.$$

⁴ An admissible solution exists by hypothesis (ii) of the theorem.

From Notes 1 and 2 it follows that

$$\begin{aligned} J(t) &= \left. \frac{\partial x}{\partial \xi} \right|_{\xi=0} \\ &= \int_{t_0}^t \left[\frac{\partial f}{\partial x}(\tau, 0, 0, 0) \frac{\partial x}{\partial \xi} + \frac{\partial f}{\partial x_d}(\tau, 0, 0, 0) \frac{\partial x_d}{\partial \xi} + \frac{\partial f}{\partial u}(\tau, 0, 0, 0) \frac{\partial u}{\partial \xi} \right]_{\xi=0} d\tau \\ &= \int_{t_0}^t \left[A(\tau)J(\tau) + B(\tau)J(\tau - h) + C(\tau) \frac{\partial u}{\partial \xi}(\tau, 0) \right] d\tau. \end{aligned}$$

Differentiating gives

$$\dot{J}(t) = A(t)J(t) + B(t)J(t - h) + C(t) \frac{\partial u}{\partial \xi}(t, 0), \quad t_0 \leq t \leq t_1 + h.$$

But from (12),

$$\frac{\partial u}{\partial \xi}(t, 0) = C'(t)K'(t, t_1), \quad t_0 \leq t \leq t_1,$$

and

$$C(t) \frac{\partial u}{\partial \xi}(t, 0) = -B(t)J(t - h), \quad t_1 < t < t_1 + h.$$

Therefore,

$$\begin{aligned} \dot{J}(t) &= A(t)J(t) + B(t)J(t - h) \\ (14) \quad &+ \begin{cases} C(t)C'(t)K'(t, t_1) & \text{for } t_0 \leq t \leq t_1, \\ -B(t)J(t - h) & \text{for } t_1 < t < t_1 + h. \end{cases} \end{aligned}$$

The solution of (14) over the interval $[t_0, t_1]$ is then

$$(15) \quad J(t) = \int_{t_0}^t K(s, t)C(s)C'(s)K'(s, t_1) ds, \quad t_0 \leq t \leq t_1.$$

By hypothesis, (15) implies that $\det J(t_1) \neq 0$. Moreover, on the interval $(t_1, t_1 + h)$, (14) is

$$(16) \quad \dot{J}(t) = A(t)J(t),$$

so that $J(t)$ is a fundamental matrix solution for (16). It follows that $\det J(t) \neq 0$ for $t \in [t_1, t_1 + h]$.

Since $J(t)$ is defined by (13), the above facts suggest that one may use an implicit function theorem to solve

$$x(t; t_0, \phi, \xi) = 0, \quad t_1 \leq t \leq t_1 + h,$$

for ξ as a function of ϕ . More precisely, consider the following theorem from Dieudonné [15].

THEOREM 6. *Let $\mathfrak{B}_1, \mathfrak{B}_2, \mathfrak{B}_3$ be Banach spaces and g a continuously differentiable map of an open subset S of $\mathfrak{B}_1 \times \mathfrak{B}_2$ into \mathfrak{B}_3 . Let $(x_0, y_0) \in S$, where $g(x_0, y_0) = 0$ and let the Fréchet derivative of g with respect to y be a linear homeomorphism of \mathfrak{B}_2 onto \mathfrak{B}_3 . Then there exists an open neighborhood N_0 of x_0 in \mathfrak{B}_1 such that for every open connected neighborhood N of x_0 contained in N_0 , there exists a unique continuous map $\Pi: N \rightarrow \mathfrak{B}_2$ such that $\Pi(x_0) = y_0$, $(x, \Pi(x)) \in S$ and $g(x, \Pi(x)) = 0$ for all $x \in N$. Furthermore, Π is continuously differentiable in N .*

Application. Let $\mathfrak{B}_1 =$ the space \mathfrak{B} of all real, n -vector-valued continuous functions on $[t_0 - h, t_0]$, $\mathfrak{B}_2 = R^n$, $\mathfrak{B}_3 =$ the space of all real, n -vector-valued continuous functions on $[t_1, t_1 + h]$, $g =$ a solution segment of (1), i.e., $g(\cdot, \cdot) = x_{[\cdot, \cdot]}(t_0, \cdot, \cdot)$. Let $S = \mathfrak{B} \times \Gamma$, where $\Gamma \subset R^n$ is an open neighborhood of the origin in R^n and represents the permissible range of ξ . (Thus $(0^{\mathfrak{B}}, 0^{\Gamma})$ is an interior point of S .) The Fréchet derivative of g with respect to ξ is a map which takes R^n into \mathfrak{B}_3 . The fact that the Jacobian matrix $J(t)$ is a homeomorphism of R^n onto R^n for each $t \in [t_1, t_1 + h]$ implies that the Fréchet derivative of g is a homeomorphism from R^n onto \mathfrak{B}_3 . Now, since $x_{[t_1, t_1 + h]}(t_0, 0^{\mathfrak{B}}, 0^{\Gamma}) = 0$, by Theorem 6 there exist a neighborhood $N(0^{\mathfrak{B}})$ of $0^{\mathfrak{B}}$ and a unique continuous map $\Pi: N(0^{\mathfrak{B}}) \rightarrow R^n$ such that $\phi \in N(0^{\mathfrak{B}})$ implies $(\phi, \Pi(\phi)) \in S$ and $x_{[t_1, t_1 + h]}(t_0, \phi, \Pi(\phi)) = 0$; that is, if $\phi \in N(0^{\mathfrak{B}})$, then

$$x(t; t_0, \phi, \xi) = 0, \quad t_1 \leq t \leq t_1 + h,$$

has an admissible solution $\xi = \Pi(\phi)$.

This completes the proof of Theorem 5.

Theorems 4 and 5 provide sufficient conditions for controllability to the origin with respect to the space \mathfrak{B} for the system (1).

6. Controllability to a function. To repeat our earlier assertion: controllability to the origin does not necessarily imply controllability to a function or to a function space. To illustrate this, and to show how the techniques presented thus far can be adapted to study controllability to a function, we present some results for controllability of (2) (and local controllability of (1)) to a function in the space \mathfrak{F} of real n -vector-valued C^1 -functions defined on the interval $[t_0 - h, t_0]$.

THEOREM 7. *Consider a pointwise complete system (2) and let $\mathfrak{L}_t(\cdot) = d(\cdot)/dt - A(t)(\cdot)$. Let $\alpha \in \mathfrak{F}$. Then (2) is controllable to $\alpha \in \mathfrak{F}$ with respect to \mathfrak{B} if and only if*

- (i) *there exists $t_1 > t_0$ such that (7) holds;*
- (ii) *with t_1 defined as above, for any $\phi \in \mathfrak{B}$, and for some admissible $u_{[t_0, t_1]}$ such that $x(t_1; t_0, \phi, u_{[t_0, t_1]}) = \alpha(t_0 - h)$, there exists an admissible solution to*

$$(17) \quad C(t)u(t) = (\mathcal{L}_t \alpha)(t - t_1 + t_0 - h) - B(t)x(t - h; t_0, \phi, u_{[t_0, t_1]})$$

on the interval $(t_1, t_1 + h)$.

Proof. The proof is essentially the same as for Theorem 1.

Now consider the following definitions.

DEFINITION 9. A system (1) is *locally controllable to a function* $\alpha \in \mathcal{FC}$ *with respect to* \mathcal{B} *if, given any initial time* t_0 *and a trajectory* $x^0(\cdot; t_0, \phi_\alpha, u_\alpha)$, $\phi_\alpha \in \mathcal{B}$, u_α *admissible, such that, for some time* $t_1 > t_0$, $x^0(t; t_0, \phi_\alpha, u_\alpha) = \alpha(t - t_1 + t_0 - h)$ *for all* $t \in [t_1, t_1 + h]$, *then there is a neighborhood* $N(\phi_\alpha)$ *of the initial function* ϕ_α *such that for each* $\phi \in N(\phi_\alpha)$ *there exists an admissible control* u^* *defined on* $[t_0, t_1 + h]$ *such that* $x(t; t_0, \phi, u^*) = \alpha(t - t_1 + t_0 - h)$ *for all* $t \in [t_1, t_1 + h]$.

DEFINITION 10. The *first variation of* (1) *about the trajectory* $x^0(\cdot; t_0, \phi_\alpha, u_\alpha)$ *is given by* (2), *where*

$$(18) \quad A(t) = \frac{\partial f}{\partial x}(t, x^0(t; t_0, \phi_\alpha, u_\alpha), x^0(t - h; t_0, \phi_\alpha, u_\alpha), u_\alpha(t)),$$

$$(19) \quad B(t) = \frac{\partial f}{\partial x_d}(t, x^0(t; t_0, \phi_\alpha, u_\alpha), x^0(t - h; t_0, \phi_\alpha, u_\alpha), u_\alpha(t)),$$

$$(20) \quad C(t) = \frac{\partial f}{\partial u}(t, x^0(t; t_0, \phi_\alpha, u_\alpha), x^0(t - h; t_0, \phi_\alpha, u_\alpha), u_\alpha(t)).$$

We then have the following theorem.

THEOREM 8. A system (1) is *locally controllable to a function* $\alpha \in \mathcal{FC}$ *with respect to* \mathcal{B} *if its first variation about the trajectory* $x(\cdot; t_0, \phi_\alpha, u_\alpha)$ *as defined in Definition 9 satisfies the conditions:*

- (i) *equation (7) holds for* t_1 *as defined in Definition 9;*
- (ii) *with* t_1 *as above,* $(\mathcal{L}_t \alpha)(t - t_1 + t_0 - h) \in \text{range } C(t)$ *almost everywhere on* $(t_1, t_1 + h)$;
- (iii) *there exists an* $n \times p$ *matrix* $D(t)$ *with measurable bounded elements such that* $B(t) = C(t)D(t)$ *almost everywhere on* $(t_1, t_1 + h)$.

Proof. The proof is essentially the same as that for Theorem 5, but it is outlined here for illustrative purposes.

Let $x^0(t; t_0, \phi_\alpha, u_\alpha) \equiv x^0(t)$ and perform the substitution in (1):

$$(21) \quad x(t) = y(t) + x^0(t).$$

Then (1) can be written as

$$(22) \quad \dot{y}(t) = -\dot{x}^0(t) + f(t, x(t), x(t - h), u(t)).$$

Solving for y assuming the zero initial function (corresponding to initial function $\phi_\alpha \in \mathcal{B}$ for x) we obtain

$$(23) \quad y(t) = -x^0(t) + \phi_\alpha(t_0) + \int_{t_0}^t f(\tau, x(\tau), x(\tau - h), u(\tau)) d\tau.$$

Now introduce a parameter ξ into $u(t)$ and let

$$(24) \quad u^\xi(t) = u(t, \xi) = \begin{cases} u_\alpha(t) + C'(t)K'(t, t_1)\xi & \text{for } t_0 \leq t \leq t_1, \\ u_\alpha(t) + \text{solution}^5 \text{ to } C(t)u(t) = -B(t) \\ \quad \cdot y(t-h; t_0, 0, u|_{t_0, t_1}^\xi) & \text{for } t_1 < t < t_1 + h, \end{cases}$$

where K represents the kernel matrix in the solution of (3) with $A(\cdot)$, $B(\cdot)$ given by (18) and (19), and $C(t)$ given by (20). Let the corresponding solution of (22) be $y(t; t_0, 0, \xi)$ and define

$$J(t) = \left. \frac{\partial y(t; t_0, 0, \xi)}{\partial \xi} \right|_{\xi=0}.$$

Since $u^0(t) = u_\alpha(t)$ and $y(t; t_0, 0, 0) = 0$, we have, upon differentiating (23),

$$J(t) = \int_{t_0}^t \left[A(\tau)J(\tau) + B(\tau)J(\tau-h) + C(\tau) \left. \frac{\partial u^\xi(\tau)}{\partial \xi} \right|_{\xi=0} \right] d\tau,$$

where $A(\cdot)$, $B(\cdot)$, $C(\cdot)$ are as in (18), (19), (20), respectively.

The remaining steps are now exactly as in the proof of Theorem 5, i.e., it follows from (7), (17) and (24) that $\det J(t) \neq 0$ for all $t \in [t_1, t_1 + h]$. Hence we can apply Theorem 6 to show existence of a solution to

$$(25) \quad y(t; t_0, \phi, \xi) = 0 \quad \text{for all } t \in (t_1, t_1 + h)$$

of the form $\xi = \Pi(\phi)$ for ϕ in some small neighborhood of the origin in y -space. (And the range of the control is contained in a neighborhood of the range of u_α .) But since, by definition,

$$\begin{aligned} y(t; t_0, \phi, \xi) &= x(t; t_0, \phi^*, \xi) - x^0(t; t_0, \phi_\alpha, u_\alpha) \\ &= x(t; t_0, \phi^*, \xi) - \alpha(t - t_1 + t_0 - h), \quad t \in (t_1, t_1 + h), \end{aligned}$$

where $\phi^* = \phi - \phi_\alpha$, the solution of (25) implies that the equation

$$x(t; t_0, \phi^*, \xi) = \alpha(t - t_1 + t_0 - h), \quad t \in (t_1, t_1 + h),$$

has a solution $\xi = \Pi^*(\phi^*)$ for all ϕ^* in a small neighborhood of ϕ_α and with the range of the control contained in a neighborhood of the range of u_α .

To obtain sufficient conditions for controllability of (1) to $\alpha \in \mathcal{H}$ with respect to \mathcal{B} , we need merely complement Theorem 8 with a theorem which yields quasi-controllability of (1) to $\alpha \in \mathcal{H}$. Such a theorem is easily obtained by rewriting Theorem 4 so that it pertains to (22).

7. Acknowledgment. The author wishes to thank Professor J. K. Hale and Professor R. K. Miller for a number of discussions which have served to increase the clarity of exposition of this paper.

⁵ An admissible solution exists by hypotheses (ii) and (iii) of the theorem.

REFERENCES

- [1] N. H. CHOKSY, *Time delay systems—a bibliography*, IRE Trans. Automatic Control, AC-5 (1960), pp. 66–70.
- [2] I. M. REPIN, *On the approximate replacement of systems with lag by ordinary dynamical systems*, Prikl. Mat. Meh., 29 (1965), pp. 226–235.
- [3] K. L. COOKE, *Functional-differential equations: Some models and perturbation problems*, Proc. Symposium on Differential Equations and Dynamical Systems, Academic Press, New York, 1967, pp. 167–183.
- [4] M. N. OĞUZTÖRELI, *Time-Lag Control Systems*, Academic Press, New York, 1966.
- [5] LEONARD WEISS AND R. E. KALMAN, *Contributions to linear system theory*, Internat. J. Engrg. Sci., 3 (1965), pp. 141–171.
- [6] L. MARKUS AND E. B. LEE, *On the existence of optimal controls*, Trans. ASME Ser. D. J. Basic Engrg., 84D (1962), pp. 13–20.
- [7] R. E. KALMAN, Discussion of [6] (see [6, pp. 21–22]).
- [8] D. H. CHYUNG AND E. B. LEE, *Optimal systems with time delays*, Proc. Third Congress of the International Federation of Automatic Control, London, 1966.
- [9] R. D. DRIVER, *Existence and stability of solutions of a delay-differential system*, Arch. Rational Mech. Anal., 10 (1962), pp. 401–426.
- [10] R. BELLMAN AND K. L. COOKE, *Differential-Difference Equations*, Academic Press, New York, 1963.
- [11] E. GODDINGTON AND R. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [12] G. BIRKHOFF AND S. MACLANE, *A Survey of Modern Algebra*, Macmillan, New York, 1953.
- [13] V. DOLEŽAL, *The existence of a continuous basis of a certain linear subspace of E , which depends on a parameter*, Casopis Pěst. Mat., 89 (1964), pp. 466–468.
- [14] N. N. KRASOVSKII, *Stability of Motion*, Stanford University Press, Stanford, 1963.
- [15] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- [16] LEONARD WEISS, *On the canonical structure of linear time-varying systems and minimal realizations of weighting patterns*, Tech. Rep. 67–5, Center for Dynamic Systems, Brown University, 1962.
- [17] LEONARD WEISS AND P. L. FALB, *On a theorem of V. Doležal with applications to system theory*, to appear.

ON THE ULTIMATE BOUNDEDNESS OF MOMENTS ASSOCIATED WITH SOLUTIONS OF STOCHASTIC DIFFERENTIAL EQUATIONS*

MOSHE ZAKAI†

1. Introduction. Let x_t be the solution to the stochastic differential equation

$$(1) \quad dx_t = m(t, x_t) dt + G(t, x_t) dw_t, \quad x_0 = a,$$

where x and $m(t, x)$ are vectors in the Euclidean r -space E , G is an $r \times q$ matrix and w_t is the standard q -dimensional Brownian motion. The following assumptions are made on m and G :

- (a) $m(t, x)$ and $G(t, x)$ are continuous in $[0, \infty) \times E$,
- (b) $|m(t, x) - m(t, y)| + |G(t, x) - G(t, y)| \leq c|x - y|$,

where for vectors, $|m| = (\sum_i m_i^2)^{1/2}$, and for matrices, $|G| = (\sum_{i,j} G_{ij}^2)^{1/2}$.

The conditional expectation $E_a V(x_t)$ will be said to be *ultimately bounded* (see [1, p. 129]) if for all a in E ,

$$(2) \quad \overline{\lim}_{t \rightarrow \infty} |E_a V(x_t)| \leq k < \infty,$$

k being independent of a . A Liapunov-type condition for the ultimate boundedness of certain functions $V(x)$ will be derived in the next section.

Stability properties of x_t in the sense of some convergence of x_t to the null state have been discussed in the literature (i.e., [2], [3], [4]); however, the class of processes having such stability properties is, for many applications, too restricted. In [5] Wonham suggested considering the property that x_t admits a stationary probability measure as a weak stability property and derived Liapunov-type conditions for stability in the weak sense. We may also consider ultimate boundedness as a form of weak stability (with respect to $V(x)$). For example, we may define x_t to be *n th order weakly stable* if x_t is ultimately bounded for $V(x) = |x|^n$.

Recently Wonham [6] derived Liapunov-type sufficient conditions for the ultimate boundedness of certain functions of x_t under the assumption that the x_t process admits an invariant probability measure. Such an assumption has been avoided in §2 (at the expense of a stronger requirement from the Liapunov function). Consequently, the results are applicable to a wider class of processes including nonstationary processes and

* Received by the editors December 12, 1966, and in revised form March 11, 1967.

† Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel.

processes for which the matrix GG' is singular, and the proofs are based on standard results on stochastic differential equations.

Two examples will be considered in §3.

2. A criterion for ultimate boundedness. Let \mathfrak{G} denote the differential operator associated with (1):

$$(3) \quad \mathfrak{G} = \sum_{i=1}^n m_i(t, x) \frac{\partial}{\partial x_i} + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n g_{i,j}(t, x) \frac{\partial^2}{\partial x_i \partial x_j},$$

where g_{ij} is the i, j th entry of GG' , where the prime denotes the transpose. We will consider functions $V(x)$ with the following properties:

- (A) $V(x)$ is real-valued, nonnegative and twice continuously differentiable in E .
- (B) Let $f(a, t)$ stand for any of the functions $E_a V(x_t)$, $E_a |\mathfrak{G}V(x_t)|$, or $E_a |(\partial V(x_t)/\partial x_i) G_{ij}(t, x_t)|^2$. Then $f(a, t)$ is, for each a , bounded in t in any bounded t interval.

It follows directly from [7] that if (1) satisfies conditions (a) and (b), then $E_a |x_t|^p$, $p > 0$, is bounded in any bounded t interval. Therefore, condition (B) is satisfied when V , $|\mathfrak{G}V|$ and $|(\partial V/\partial x_i)G_{ij}|$ are dominated by polynomials.

THEOREM 1. *If x_t satisfies (1) with (a) and (b) and if $V(x)$ satisfies (A) and (B), then*

$$(4) \quad \mathfrak{G}V(x) \leq k_1 - k_2 V(x), \quad k_2 > 0, \quad k_1 \geq 0,$$

for all $t > 0$ implies

$$(5) \quad E_a V(x_t) \leq V(a)e^{-k_2 t} + \frac{k_1}{k_2} (1 - e^{-k_2 t}).$$

If, in addition, m and G are independent of t and for some $\epsilon > 0$, $h > 0$,

$$(6) \quad E_a |\mathfrak{G}V(x_t)|^{1+\epsilon} \leq c < \infty$$

(c may depend on a) for all $0 \leq t \leq h$, then (5) implies (4). (The last inequality can be replaced by the more general condition that $E_a \mathfrak{G}V(x_t)$ is continuous in t at $t = 0$.)

Proof. By assumption (A), we may apply Itô's formula [8]; therefore,

$$V(x_t) - V(a) = \int_0^t \mathfrak{G}V(x_s) ds + \int_0^t \left(\frac{\partial V(x_s)}{\partial x} \right)' G(x_s) dw_s.$$

By assumption (B), the expectation of the stochastic integral is zero and

$$(7) \quad E_a V(x_t) = V(a) + \int_0^t E_a \mathfrak{G}V(x_s) ds.$$

It follows from (7) that $E_a V(x_t)$ is absolutely continuous in t (with respect to Lebesgue measure). Therefore, for almost all s , $s \geq 0$,

$$(8) \quad \frac{dE_a V(x_s)}{ds} = E_a \mathfrak{G}V(x_s)$$

$$(9) \quad \leq k_1 - k_2 E_a V(x_s).$$

For all those s for which (9) holds, we have

$$\frac{de^{k_2 s} E_a V(x_s)}{ds} \leq k_1 e^{k_2 s}.$$

Since $e^{k_2 t} E_a V(x_t)$ is also absolutely continuous, $e^{k_2 t} E_a V(x_t)$ is (everywhere) the indefinite integral of its almost-everywhere derivative. Integrating the last inequality we get

$$e^{k_2 t} E_a V(x_t) - V(a) \leq \frac{k_1}{k_2} (e^{k_2 t} - 1),$$

from which (5) follows. Conversely, if (5) is true, then for $t > 0$,

$$\frac{E_x V(x_t) - V(x)}{t} \leq \left(\frac{k_1}{k_2} - V(x) \right) (1 - e^{-k_2 t}) \frac{1}{t}$$

and

$$(10) \quad \lim_{t \rightarrow 0} \frac{E_x V(x_t) - V(x)}{t} \leq k_1 - k_2 V(x)$$

since

$$(11) \quad \frac{E_x V(x_t) - V(x)}{t} = \frac{1}{t} \int_0^t E_x \mathfrak{G}V(x_s) ds.$$

Then if $E_x \mathfrak{G}V(x_s)$ is continuous in s at $s = 0$, (4) follows from (10) and (11). In particular, since $\mathfrak{G}V(x_s)$ is a.s. continuous, (6) implies the continuity of $E_x \mathfrak{G}V(x_s)$ (see [9, Corollary 2, p. 164]).

3. Applications.

Example 1. Consider the stochastic differential equation

$$(12) \quad dx_t = Ax_t dt + f(x_t) dt + G(x_t) dw_t,$$

where A is a constant $r \times r$ matrix and $f(x)$ a vector-valued function. It is assumed that condition (b) is satisfied.

THEOREM 2. *If all the characteristic values of A have negative real parts, $|f(x)| = o(|x|)$ as $x \rightarrow \infty$ and $|G_{ij}(x)| = o(|x|)$ as $x \rightarrow \infty$, then $E_a |x_t|^p$ is ultimately bounded for every positive p .*

Proof. Since $E^{1/p} |x|^p$ is nondecreasing in p , it suffices to prove the theorem for even integers. Consider the matrix equation for B :

$$(13) \quad A'B + BA = -C,$$

where C is an arbitrary positive definite matrix. From the assumption on the characteristic values of A it follows that (13) has a positive definite solution B (see [1, p. 26]). Let λ_m and λ_M be the smallest and largest eigenvalues of B , respectively. Setting $V(x) = x'Bx$, we have $\lambda_m |x|^2 \leq V(x) \leq \lambda_M |x|^2$. For this choice of $V(x)$, we have

$$\begin{aligned} \mathfrak{G}V(x) &= x'A'Bx + x'BAx + 2x'Bf(x) + \text{trace}\{BG(x)G'(x)\} \\ &= -x'Cx + 2x'Bf(x) + \text{trace}\{BG(x)G'(x)\} \\ &= -\frac{1}{2}x'Cx + [2x'Bf(x) + \text{trace}\{BG(x)G'(x)\} - \frac{1}{2}x'Cx]. \end{aligned}$$

Since C is positive definite and $|f(x)|$ and $|G(x)|$ are $o(|x|)$ as $|x| \rightarrow \infty$, there exists an $R > 0$ such that the term in brackets will be negative for all $|x| \geq R$. Therefore, there exists a constant $c_1 > 0$ such that for all x in E ,

$$(14) \quad \begin{aligned} \mathfrak{G}V(x) &\leq -\frac{1}{2}x'Cx + c_1 \\ &\leq -\frac{1}{2}\frac{\lambda_c}{\lambda_M}V(x) + c_1, \end{aligned}$$

where λ_c is the smallest eigenvalue of C . Since the last inequality satisfies Theorem 1, $E_a |x_t|^2 \leq \lambda_m^{-1} E_a V(x_t)$ is ultimately bounded. For $p = 2n$,

$$\mathfrak{G}V^n(x) = nV^{n-1}(x)\mathfrak{G}V(x) + n(n-1)V^{n-2}(x)\text{trace}\left\{\left(\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j}\right)G(x)G'(x)\right\},$$

and by (14),

$$\begin{aligned} \mathfrak{G}V^n(x) &\leq -\frac{1}{4}\frac{\lambda_c}{\lambda_M}nV^n(x) + \left[-\frac{1}{4}\frac{\lambda_c}{\lambda_M}nV^n(x) \right. \\ &\quad \left. + n(n-1)V^{n-2}(x)\text{trace}\left\{\left(\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j}\right)G(x)G'(x)\right\} + nc_1V^{n-1}(x)\right]. \end{aligned}$$

Since $V^n(x)$ is of the order of $|x|^{2n}$ and the other terms in the square bracket are of lower order, there exist constants $c_2, c_3 > 0$ such that for all x in E ,

$$\mathfrak{G}V^n(x) \leq -c_2V^n(x) + c_3.$$

The ultimate boundedness of $E_a |x_t|^{2n}$ follows from Theorem 1 and $E_a |x_t|^{2n} \leq \lambda_m^{-n} E_a V^n(x_t)$.

Example 2. A generalization of Example 1 in [5] will be considered here; in particular, G will not be required to be strictly positive definite. Let

$$(15) \quad \begin{aligned} dx_i &= Fx_i dt - b\phi(\sigma) dt + f(x_i) dt + G(x_i) dw_i, \\ \sigma &= c'x, \end{aligned}$$

where F is a constant $r \times r$ matrix, b and c are constant r -vectors, ϕ is scalar-valued and f is vector-valued.

THEOREM 3. *Let the system (15) satisfy (b) and the following conditions:*

- (i) *All the eigenvalues of F have negative real parts.*
- (ii) *$\sigma\phi(\sigma) > 0$ for all $\sigma \neq 0$, $\phi(\sigma)$ is continuously differentiable and $\dot{\phi}(\sigma) = d\phi/d\sigma$ is bounded in $(-\infty, \infty)$.*
- (iii) *$|f(x)| = o(|x|)$ and $|G(x)| = o(|x|)$ as $|x| \rightarrow \infty$.*
- (iv) *There exist two nonnegative constants α and β such that $\alpha + \beta > 0$,*

$$\operatorname{Re}(\alpha + i\omega\beta)c'(i\omega I - F)^{-1}b > 0$$

for all real ω , and if $\beta c'b = 0$, then $\alpha c'Fb + \beta c'F^2b \neq 0$.

Then $E_a |x_t|^p$ is ultimately bounded for every positive p .

Proof. For $p = 2$, set

$$V(x) = x' Bx + \beta \int_0^{c'x} \phi(\sigma) d\sigma.$$

Then $\operatorname{grad} V(x) = 2Bx + \beta\phi(c'x)c$ and

$$\left(\frac{\partial^2 V}{\partial x_i \partial x_j} \right) = 2B + \beta(\dot{\phi})(c'x)cc',$$

which is bounded. Therefore,

$$\mathfrak{G}V(x) \leq o(|x|^2) + (\operatorname{grad} V(x))'(Fx - b\phi(\sigma)).$$

By the results of Yacubovich-Kalman-Meyer [10, Lemma 4 and Theorem 1], there exist positive definite matrices B and Q such that

$$\mathfrak{G}V(x) \leq o(|x|^2) - x'Qx.$$

Since Q is positive definite and $|\phi(\sigma)| \leq \lambda|\sigma|$, $\beta \geq 0$, there exists $\mu > 0$ such that $\mu V(x) \leq x'Qx$. Therefore, there exist k_1 and k_2 such that Theorem 1 applies for $V(x)$, and since $V(x) \geq c|x|^2$, $E_a |x_t|^2$ is ultimately bounded. For any integer $n > 1$, we have

$$\begin{aligned} \operatorname{grad} V^n(x) &= nV^{n-1}(x) \operatorname{grad} V(x), \\ \left(\frac{\partial^2 V^n(x)}{\partial x_i \partial x_j} \right) &= nV^{n-1}(x)(2B + \beta\dot{\phi}(c'x)cc') \\ &\quad + n(n-1)V^{n-2}(x)(\operatorname{grad} V(x))(\operatorname{grad} V(x))'. \end{aligned}$$

Therefore,

$$\mathfrak{G}V^n(x) \leq o(|x|^2) - n\mu V^n(x)$$

and $E_a V^n(x_t)$, $E_a |x_t|^{2n}$ are ultimately bounded.

Acknowledgment. The author wishes to thank a member of the editorial board for pointing out an error in the previous version of this paper and for calling his attention to the correction note of [10].

REFERENCES

- [1] W. HAHN, *Theory and Application of Liapunov's Direct Method*, Prentice Hall, Englewood Cliffs, New Jersey, 1963.
- [2] R. Z. KHASMINSKII, *On the stability of the trajectory of Markov processes*, J. Appl. Math. Mech., 26 (1963), pp. 1554-1565.
- [3] H. J. KUSHNER, *On the theory of stochastic stability*, Tech. Rep. 65-1, Center for Dynamical Systems, Brown University, Providence, Rhode Island, 1965. Also in *Advances in Control Systems*, vol. 4, C. T. Leondes, ed., Academic Press, New York, 1967.
- [4] F. KOZIN, *On almost sure asymptotic sample properties of diffusion processes defined by stochastic differential equations*, J. Math. Kyoto Univ., 4 (1965), pp. 515-528.
- [5] W. M. WONHAM, *Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2 (1966), pp. 195-207.
- [6] ———, *A Liapunov method for the estimation of statistical averages*, Ibid., to appear.
- [7] M. ZAKAI, *Some moment inequalities for stochastic integrals and solutions to stochastic differential equations*, Israel J. Math., to appear.
- [8] K. ITÔ, *On a formula concerning stochastic differentials*, Nagoya Math. J., 3 (1951), pp. 55-65.
- [9] M. LOÈVE, *Probability Theory*, Van Nostrand, Princeton, New Jersey, 1963.
- [10] K. R. MEYER, *On the existence of Lyapunov functions for the problem of Luré*, this Journal, 3 (1966), pp. 373-383; note of correction, J. Differential Equations, to appear.

LAGRANGIAN SADDLE POINTS AND OPTIMAL CONTROL*

D. O. NORRIS†

1. Introduction. In [1] Hurwicz has treated the relationship between a Lagrangian saddle point and the maximum of a constrained function. The theorem [1, Theorem V.3.1], which gives conditions under which the existence of a maximum implies the existence of a saddle point, requires conditions (viz., positive cones have nonempty interior) which are often too stringent for application. In [2] a weakening of these conditions was obtained (see [2, Theorem 1]); however, these results are still not applicable for some control problems.

For example, consider the following well-known problem (see [5, pp. 475-480]). Suppose the performance of a system is described by the differential equation

$$(1) \quad x' = Ax + Bu,$$

where A is $n \times n$, B is $n \times r$, and the components of these matrices are bounded measurable functions. The control u is required to belong to the set $U = \{u: u = (u_1, \dots, u_r), |u_j(t)| \leq 1, u_j(t) \in L_\infty, j = 1, \dots, r\}$. Given an initial vector $x_0 \in R^n$ and a terminal time T , the problem is to choose $u \in U$ such that u transfers x_0 to the origin $0 \in R^n$ in $[0, T]$ and such that

$$(2) \quad J(u) = \sum_{j=1}^r \int_0^T u_j^2(t) dt$$

is minimized. In view of the form of J it seems desirable to consider the problem in $L_2^r [0, T]$, the set of all r -tuples of functions (as an abuse of language we identify a function and its equivalence class) in $L_2 [0, T]$, where

$$(3) \quad \|u\| = \left[\sum_{j=1}^r \int_0^T u_j^2(t) dt \right]^{1/2}.$$

$L_2^r [0, T]$ is a B -space with this norm. Let $H: L_2^r [0, T] \rightarrow R^n$ such that

$$(4) \quad H(u) = \int_0^T \Phi^{-1}(s)B(s)u(s) ds + x_0,$$

where Φ is the principal matrix solution of $x' = Ax$. $H(u) = 0$ if and only if u transfers x_0 to 0 in $[0, T]$. In addition, let $G: L_2^r [0, T] \rightarrow L_2^r [0, T]$ be

* Received by the editors March 24, 1967, and in revised form July 21, 1967.

† Air Force Institute of Technology, Wright-Patterson Air Force Base, Dayton, Ohio 45433.

such that

$$(5) \quad G(u) = \begin{bmatrix} u_1^2 - 1 \\ \vdots \\ u_r^2 - 1 \end{bmatrix}.$$

The optimal control problem can now be recast as a typical programming problem. Namely, minimize J subject to the constraints $u \in U$, $H(u) = 0$, and $G(u) \leq 0$. Unfortunately, the positive cone in $L_2^r[0, T]$ (i.e., $u \geq 0$ if and only if $u_j(t) \geq 0$ a.e. on $[0, T]$, $j = 1, \dots, r$) has an empty interior, so Theorem V.3.1 of [1] does not apply. Furthermore, it is not clear that the generalizations obtained in [2] apply either.

The main result of this paper is the proof of a theorem which demonstrates that a necessary condition for the existence of a constrained minimum is the existence of a saddle point under conditions which are applicable to the problem stated above. In fact, condition (N''') in [2, Theorem I] will be omitted.

2. The main theorem. In this section the necessary condition for a constrained minimum will be presented. Chapter 5 of [3] should be consulted for a detailed account of order properties used here and in the next section. In this paper the value of a linear functional y' at a point y will be denoted by $y'y$ or $y'(y)$.

THEOREM 1. *X denotes a real vector space and K is convex in X . Y denotes a normed linear space which is an ordered vector space. $J: X \rightarrow R^1$ and $G: X \rightarrow Y$ are convex on K .*

(R_c'') *For each positive, nonnull, continuous linear functional $y^* \in Y^*$, the topological dual of Y , there is an $x \in K$ such that $y^*G(x) < 0$.*

If $J(x_0) \leq J(x)$ on K subject to the constraint $G(x) \leq 0$, then there are a real number $\eta_0 > 0$ and $y_0^ \in Y^*$ with $y_0^* \geq 0$ such that the Lagrangian $\phi(x, y^*; \eta_0) = \eta_0 J(x) + y^*G(x)$ has a saddle point at $(x_0, y_0^*; \eta_0)$ for every $x \in K$, $y^* \geq 0$.*

2.1. Discussion. If C is the positive cone in Y , C is closed, Y is complete, and $Y = C - C$, then every positive linear functional on Y is continuous (see [3, Theorem 5.5]), so in this case it suffices to look at positive linear functionals. If Y is an order complete vector lattice of minimal type and K contains a weak order unit x (see [3, pp. 241–242]) such that $G(x) \leq 0$, $G(x) \neq 0$, then every nonnull positive linear functional y' on Y has the property that $y'(x) < 0$. Since an order complete vector lattice Y has the property that $Y = C - C$, a sufficient condition for (R_c'') to hold is that Y be an ordered topological vector space which is an order complete vector lattice of minimal type such that K contains a weak order unit x with $G(x) \leq 0$, $G(x) \neq 0$.

The condition that Y be a normed space cannot be weakened since the proof of the theorem requires $W = R^1 \times Y$ to be a normed space. The condition that the range of J be in R^1 is not essential. However, the range of J must be a normed space in which the following property holds: If $\beta \leq J(x_0)$ then $\inf_{\alpha \geq J(x_0)} \|\beta - \alpha\| = \|\beta - J(x_0)\|$.

Finally, the theorem is stated as a minimization theorem with convex J and G , since this is the type of problem presented in §1. The problem treated in [1] and [2] is a maximization problem for concave functions. It should be noted that the saddle point inequalities are reversed for a minimization problem, i.e.,

$$(6) \quad \phi(x, y_0^*; \eta_0) \geq \phi(x_0, y_0^*; \eta_0) \geq \phi(x_0, y^*; \eta_0)$$

for all $x \in K, y^* \geq 0$.

2.2. Proof of the theorem. Let $W = R^1 \times Y$, where if $w = (\alpha, y) \in W$ then $\|w\| = |\alpha| + \|y\|$. For each $x \in K$ define $A(x) = \{(\alpha, y) : \alpha \geq J(x), y \geq G(x)\}$ and let $A = \bigcup_{x \in K} A(x)$. A is convex because J and G are convex functions on the convex set K . $(J(x_0), 0) \in A$ since $x_0 \in K$ and $0 \geq G(x_0)$.

A can be supported at $(J(x_0), 0)$ by a nonzero continuous linear functional. It is sufficient [4, Theorem 11, p. 452] to show that, for some point $w_1 \notin A$,

$$(7) \quad \|w_1 - (J(x_0), 0)\| = \inf_{w \in A} \|w_1 - w\|.$$

If $\beta < J(x_0)$ then $w_1 = (\beta, 0)$ will satisfy (7). Let w_0^* be the supporting functional. Then $w_0^*(w) \geq w_0^*(J(x_0), 0)$ for every $w \in A$. Now $w_0^* = (\eta_0, y_0^*)$ for some real number η_0 and $y_0^* \in Y^*$.

It is now necessary to show that w_0^* is positive, $\eta_0 > 0$ and the saddle point inequalities are satisfied. The verification of these facts is essentially the same as they appear in [1] and [2] but will be included here for completeness. If $w = (\alpha, y) \in A$, then

$$(8) \quad w_0^*(w) = \eta_0 \alpha + y_0^* y \geq \eta_0 J(x_0) = w_0^*(J(x_0), 0).$$

$(\alpha, 0) \in A$ for every $\alpha \geq J(x_0)$, so from (8) it follows that $\eta_0 \alpha \geq \eta_0 J(x_0)$ and, consequently, $\eta_0 \geq 0$. $(J(x_0), y) \in A$ for every $y \geq 0$ so again from (8) it follows that $y_0^* y \geq 0$ for every $y \geq 0$ and, consequently, $y_0^* \geq 0$. Now we show $\eta_0 > 0$. For, suppose not, then since $w_0^* \neq 0$ it follows that $y_0^* \geq 0, y_0^* \neq 0$, and from (8) with $\eta_0 = 0$ we conclude that $y_0^* y \geq 0$ for every $y \geq G(x), x \in K$. In particular, it holds for every $G(x)$ and we have $y_0^* G(x) \geq 0$. But $G(x) \leq 0, y_0^* \geq 0$ so $y_0^* G(x) = 0$ for every $x \in K$ such that $G(x) \leq 0$. This contradicts the hypothesis of the theorem. The saddle point inequalities are now easily verified. If $w = (J(x_0), G(x_0))$, then from

(8), $y_0^* G(x_0) \geq 0$, and since $y_0^* G(x_0) \leq 0$, we conclude $y_0^* G(x_0) = 0$. Furthermore, $y^* G(x_0) \leq 0$ for every $y^* \geq 0$, so $\eta_0 J(x_0) + y_0^* G(x_0) \geq \eta_0 J(x_0) + y^* G(x_0)$. Again from (8), since $(J(x), G(x)) \in A$ for every $x \in K$, $\eta_0 J(x) + y_0^* G(x) \geq \eta_0 J(x_0) + y_0^* G(x_0)$, and the proof is now complete.

It should be noted that the sufficient conditions given in [1, Theorem V.1] for the existence of a minimum in terms of a saddle point still apply with the obvious modifications needed to change a maximization problem into a minimization problem. The necessary and sufficient conditions given in Theorem 1 above and in [1, Theorem V.1] will now be applied to the control problem specified in §1.

3. Application. Let $X = Y = L_2^r[0, T]$. The positive cone C in Y is $C = \{u: u_j \geq 0 \text{ a.e. on } [0, T], j = 1, \dots, r\}$. Let $K = \{u \in X: H(u) = 0\}$, where H is defined in (4). K is convex. $L_2^r[0, T]$ is an ordered topological vector space which is an order complete vector lattice of minimal type. So to apply the theorem it is necessary to show that for each nonnull, positive $y^* \in Y^*$ there is a $u \in K$ such that $y^* G(u) < 0$, $G(u) \leq 0$. It will suffice to show that there is a $u \in K$ such that $G(u) < 0$ (i.e., each component $u_j^2(t) - 1 < 0$ a.e. on $[0, T]$). In this case, $G(u)$ will be the negative of a weak order unit and consequently, every nonnull positive $y^* \in Y^*$ will be negative on $G(u)$. Let T^* denote the minimum time for which transfer can be effected using a "bang-bang" control (i.e., $G(u) = 0$). If $T^* > T$ then the problem has no solution. If $T^* = T$ then $J(u) = rT$ for every $u \in K$ and the optimum is found. If $T^* < T$, then a "bang-bang" control u with amplitude strictly less than one can be used to transfer x_0 to 0 in $[0, T']$, where $T^* \leq T' \leq T$. This is a consequence of the fact that H is continuous. For this control, $G(u) < 0$. Thus, we have established that the conditions of the theorem are satisfied. These are also adequate for the sufficiency.

It will now be shown that necessary and sufficient conditions for u_0 to minimize J on K such that $G(u) \leq 0$ are that there exist y_0^* represented in $L_2^r[0, T]$ by $(\lambda_1^*, \dots, \lambda_r^*)$, $\eta_0 > 0$, and $\eta^* \in R^n$ such that

$$(9) \quad \lambda_j^* \geq 0, \quad j = 1, \dots, r,$$

$$(10) \quad \lambda_j^* [u_{0j}^2 - 1] = 0, \quad j = 1, \dots, r,$$

$$(11) \quad \langle \eta_0 u_0 + [\lambda^* u_0], h \rangle = 0$$

for all $h \in L_2^r[0, T]$, such that $H(h) - x_0 = 0$,

$$\langle u, v \rangle = \sum_{j=1}^r \int_0^T u_j(t) v_j(t) dt,$$

and $[\lambda u] = (\lambda_1 u_1, \dots, \lambda_r u_r)$.

The Lagrangian $\phi(u, y^*; \eta_0)$ is Fréchet differentiable in u . In fact,

$$(12) \quad \begin{aligned} \phi(u + h, y^*; \eta_0) - \phi(u, y^*; \eta_0) \\ = 2 \langle \eta_0 u + [\lambda u], h \rangle + \eta_0 \|h\|^2 + \langle [\lambda h], h \rangle, \end{aligned}$$

where y^* is represented by $(\lambda_1, \dots, \lambda_r)$. Suppose u_0 minimizes J on K such that $G(u) \leq 0$; then by Theorem 1 there are $\eta_0 > 0$ and $y_0^* \geq 0$ such that the saddle point inequalities hold for all $u \in K$ and $y^* \geq 0$. Now $y_0^* \geq 0$ if and only if $\lambda_j^* \geq 0$ a.e. on $[0, T]$, $j = 1, \dots, r$. Thus, (9) holds. From the second inequality in (6) it follows that $y_0^* G(u_0) \geq y^* G(u_0)$ for all $y^* \geq 0$. Hence, $y_0^* G(u_0) = 0$. Thus, $\sum_{j=1}^r \int_0^T \lambda_j^*(t) [u_{0j}(t)^2 - 1] dt = 0$ and it follows that $\lambda_j^*(t) [u_{0j}(t)^2 - 1] = 0$ a.e. on $[0, T]$, $j = 1, \dots, r$, since $\lambda_j^* \geq 0$ and $u_{0j}^2 - 1 \leq 0$. Thus, (10) holds. From the first inequality in (6) we have that $\phi(u_0 + h, y_0^*; \eta_0) - \phi(u_0, y_0^*; \eta_0) \geq 0$ for every h such that $(u_0 + h) \in K$. Now $(u_0 + h) \in K$ if and only if $\int_0^T Y(s)h(s) ds \equiv \langle Y, h \rangle = 0$, where $Y(s) = \Phi^{-1}(s)B(s)$. Let $Z = \{h: \langle Y, h \rangle = 0\}$. Then on Z , (12) holds and, consequently, $\langle \eta_0 u_0 + [\lambda^* u_0], h \rangle = 0$ on Z , for if not, then there is an $h_1 \in Z$ such that $\langle \eta_0 u_0 + [\lambda^* u_0], h_1 \rangle = M \neq 0$. From (6) and (12),

$$\begin{aligned} 2 \langle \eta_0 u_0 + [\lambda^* u_0], \alpha h_1 \rangle + \eta_0 \|\alpha h_1\|^2 + \langle [\lambda^* (\alpha h_1)], \alpha h_1 \rangle \\ = 2 \alpha M + \alpha^2 \eta_0 \|h_1\|^2 + \alpha^2 \langle [\lambda^* h_1], h_1 \rangle \geq 0. \end{aligned}$$

If $M > 0$, use $\alpha < 0$ and note that $2M + \alpha \eta_0 \|h_1\|^2 + \alpha \langle [\lambda^* h_1], h_1 \rangle < 0$ for all $\alpha < 0$, which is impossible since $M > 0$. If $M < 0$, use $\alpha > 0$. Thus, (11) is verified.

Conversely, if (9), (10) and (11) hold, then ϕ has a saddle point at $(u_0, y_0^*; \eta_0)$ since the nonlinear terms in h in (12) are nonnegative and, consequently, u_0 minimizes J on K such that $G(u) \leq 0$.

From (9), (10) and (11) we can deduce the necessary and sufficient conditions for optimality. Let y_1, \dots, y_n denote the rows of Y and suppose they are linearly independent as elements in $L_2^r[0, T]$. If we now think of y_1, \dots, y_n as being representations for continuous linear functionals defined on $L_2^r[0, T]$, we can define

$$Z_i = \{y \in L_2^r[0, T]: \langle y_i, y \rangle = 0\}.$$

$Z = \bigcap_{i=1}^n Z_i$ and since $\langle \eta_0 u_0 + [\lambda^* u_0], z \rangle = 0$ for all $z \in Z$, we conclude that $\eta_0 u_0 + [\lambda^* u_0]$ is a linear combination of y_1, \dots, y_n , say, $\eta_0 u_0 + [\lambda^* u_0] = \eta^* Y$. If $\lambda_j^*(t) = 0$, then

$$u_{0j}(t) = \frac{1}{\eta_0} \sum_{i=1}^n \eta_i^* y_{ij}(t)$$

and we must have $|u_{0j}(t)| < 1$. If $|u_{0j}(t)| = 1$, then

$$\lambda_j^*(t) = \pm \sum_{i=1}^n \eta_i^* y_{ij}(t) \mp \eta_0,$$

where η_0 must be chosen so that $\lambda_j^*(t)$ is nonnegative. Actually, the value of η_0 depends upon T . The closer T gets to the time T^* for the time optimal control problem, the smaller η_0 must be. Thus, the necessary and sufficient conditions for optimality imply that

$$u_{0j}(t) = \begin{cases} \operatorname{sgn} \frac{1}{\eta_0} \sum_{i=1}^n \eta_i^* y_{ij}(t) & \text{if } \left| \frac{1}{\eta_0} \sum_{i=1}^n \eta_i^* y_{ij}(t) \right| \geq 1, \\ \frac{1}{\eta_0} \sum_{i=1}^n \eta_i^* y_{ij}(t) & \text{otherwise.} \end{cases}$$

This is the familiar saturation control for the problem specified (see [5]).

REFERENCES

- [1] L. HURWICZ, *Programming in linear spaces*, Studies in Linear and Non-Linear Programming, K. J. Arrow, L. Hurwicz, H. Uzawa, eds., Stanford University Press, Stanford, 1958, pp. 38-102.
- [2] L. HURWICZ AND H. UZAWA, *A note on the Lagrangian saddle-points*, Ibid., pp. 103-113.
- [3] H. H. SCHAEFFER, *Topological Vector Spaces*, Macmillan, New York, 1966.
- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.
- [5] M. ATHANS AND P. FALB, *Modern Control Theory*, McGraw-Hill, New York, 1964.

EXTREMAL STATES AND EXTREMAL CONTROLS*

G. P. AKILOV, L. V. KANTOROVICH AND G. SH. RUBINSHTEIN†

Abstract. The article introduces the notion of an extremal state of a system which consists of a vector space, a topology in the space, a set of cones and a state set. The state is a point in this state set such that each ray from this point in a direction contained within a certain cone contains no points of the state set near the original point. Sufficient conditions for the existence of extremal states and extremality criteria are given. Possible applications of this notion are described, in particular, in linear and convex programming, in continuous transport problems, in the theory of economic models of supply, etc.

Questions on the finding and on the investigation of the solutions of extremal problems: optimal solutions, optimal trajectories, optimal controls, etc., occupy ever greater space in mathematics and its applications. In this connection in many cases we are not interested so much in the statement and solution of an actual extremal problem as we are in the general nature of the possible solutions of a certain class of extremal problems and in their properties (for instance, extremals in the calculus of variations). Often, the formulation of an actual problem as an extremum problem is to some degree artificial and conditional, whereas classes of extremal solutions arise completely naturally and the solutions in these classes are of a simple and meaningful nature.

In this article we introduce the notion of an extremal state, which is defined as the "most preferred" (locally or globally) position of a point in a given set of "admissible" states in a linear topological space. The usual extremum is a special case when the "preference" is determined by the value of a certain functional.

The article gives existence conditions, criteria and characteristics of extremal states, which generalize the known statements relative to extremals and stationary points. Also, certain conjectures are made concerning the set of extremal states, which generalize the principle of feasible displacements. Along with the notion of an extremal state we introduce the notion of an extremal control.

Certain of the numerous practical realizations possible and their applications are given: to the theory of optimal mathematical-economic models of production and consumption in linear, convex and continuous program-

* Originally published in *Vestnik Leningradskogo Universiteta*, no. 7 (1967), pp. 30-37. Submitted in January, 1967. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Leningrad University, Leningrad, USSR.

ming schemes. The possibility of analogous treatment of problems on relative extrema in classical analysis and of certain questions in the calculus of variations should also be noted.

In what follows it is possible to find a number of other generalizations and applications of this same plan.

1. We consider a real vector space Z and a Hausdorff topology τ on it, invariant relative to translation and such that there exists a fundamental system of neighborhoods of zero, consisting of balanced absorbing sets.

Let $\Omega_0 \subset Z$. By \mathfrak{M} we denote a family $\{M_z\}$, $z \in \Omega_0$, of cones (with vertex at the origin) in the space Z . We shall speak of an element $l \in M_z$ as a desirable direction at the point z .

If the nonempty set $\Omega \subset \Omega_0$, then the quadruple $[Z, \tau, \mathfrak{M}, \Omega]$ is called the *optimizable system*.

We shall call an element $z_0 \in \Omega$ an *extremal state* of the system $[Z, \tau, \mathfrak{M}, \Omega]$ if there exists a neighborhood V of the point z_0 such that the intersection $\Omega \cap V \cap (z_0 + M_{z_0})$ reduces to z_0 .

Various propositions may be proved concerning the existence of extremal states, for instance, the following theorem.

THEOREM 1. *Let the optimizable system $[Z, \tau, \mathfrak{M}, \Omega]$ satisfy the conditions:*

- (i) Ω is compact in the topology τ ;
- (ii) there exists a convex closed cone $K \subset Z$ such that

$$K \cap (-K) = \{0\} \quad \text{and} \quad \bigcup_{z \in \Omega} M_z \subset K.$$

Then, the given system has at least one extremal state.

Proof. In Ω we introduce the relation \leq by setting $z_1 \leq z_2$ ($z_1, z_2 \in \Omega$) if $z_2 - z_1 \in K$. It is clear that this relation is an ordering. We can convince ourselves that it is inductive.

Let E be a completely ordered subset of set Ω . For every $z \in E$, the set $F_z = \{u \in \Omega: z \leq u\}$ is not empty ($z \in F_z$) and, since $F_z = \Omega \cap (z + K)$, it is closed and, consequently, also compact. The family $\{F_z\}$, $z \in E$, is centered: if $z_1, z_2, \dots, z_n \in E$, then by taking $z_1 \leq z_2 \leq \dots \leq z_n$ we have $\bigcap_{k=1}^n F_{z_k} = F_{z_n} \neq \emptyset$. Hence we conclude that $\bigcap_{z \in E} F_z \neq \emptyset$. Having chosen an element $u \in \bigcap_{z \in E} F_z$, we shall have $z \leq u$ for any $z \in E$.

By Zorn's lemma a maximal element z_0 exists in Ω . This means that the intersection $\Omega \cap (z_0 + K)$ consists of a unique element z_0 . Since $M_{z_0} \subset K$, z_0 is an extremal state.

Remark 1.1. In Theorem 1, condition (ii) can be replaced by the requirement: (ii') there exists a linear continuous (in the topology τ) functional f on Z , strictly positive on every cone M_z , i.e., such that $f(u) > 0$ holds for any $z \in \Omega$ and $u \in M_z$ ($u \neq 0$).

Indeed, it suffices to note that the point z_0 where $f(z_0) = \sup f[\Omega]$ will be an extremal state.

2. We shall take the topology τ to be locally convex for the optimizable system $[Z, \tau, \mathfrak{M}, \Omega]$ considered in this section.

THEOREM 2. *Let z_0 be an extremal state of the system $[Z, \tau, \mathfrak{M}, \Omega]$ such that the cone M_{z_0} is convex and does not reduce to the null element, and let there exist a convex neighborhood V of the point z_0 for which the intersection $\Omega \cap V$ is convex. Suppose, further, one of the following conditions is satisfied:*

- (a) *the sets $\Omega \cap V$ and $(z_0 + M_{z_0}) \cap V$ are compact;*
- (b) *the set $\Omega \cap V$ has a nonempty interior;*
- (c) *the cone M_{z_0} has a nonempty interior.*

Then we can find a nontrivial linear functional f on Z , continuous in the topology τ , such that f has a local maximum on Ω at the point z_0 . Here, for any direction l desirable at the point z_0 , we have $f(l) \geq 0$.

Proof. The proof is obtained in an obvious manner by the use of well-known theorems on the separability of convex sets.

Remark 2.1. In the case when condition (c) is realized, we can assert that $f(l) > 0$ for any interior element l of cone M_{z_0} . In particular, if the set $M_{z_0} \setminus \{0\}$ is open, then $f(l) > 0$ for any nonzero element $l \in M_{z_0}$.

Remark 2.2. If the sets $V \cap \Omega$ and M_{z_0} are polyhedral, i.e., if each of them is the intersection of a finite number of closed subspaces and of a closed affine manifold, then in this case the functional f can be chosen so that $f(l) > 0$ for any $l \in M_{z_0}$ ($l \neq 0$).

If under these assumptions the extremal state $z_0 \in \Omega$ is an interior point of a face, whose defect we shall take for definiteness to be unity, and if the cones M_z do not depend on z , then, clearly, any other point z of the indicated face will also be an extremal state. In such case the theorem will be true for every such point with one and the same functional f . This means that a point z lying in a sufficiently small neighborhood of the point z_0 is an extremal state if and only if $f(z - z_0) = 0$. The stated relation can be considered as the equation for the variation of the extremal state.

If the closed space Z is finite-dimensional and if Ω is a closed convex body with a smooth boundary close to the point z_0 , then, although what we have said above in reference to the polyhedral sets no longer holds in the strict sense, nevertheless, in this case also, small variations in the hyperplane $\{u \in Z: f(u) = f(z_0)\}$ lead to points whose differences from certain extremal states are of a "high order of smallness." Here we shall not make the meaning of this statement more precise, but shall remark only that it can be made even in the case when the cones M_z do depend on z , this dependence being continuous in a specific sense.

Remark 2.3. If the point $z_0 \in \Omega$ is such that there exists a linear continuous

functional f on Z which has a local maximum on Ω at the point z_0 and if $f(l) > 0$ for any $l \in M_{z_0}$ ($l \neq 0$), then z_0 is an extremal state.

It follows from what we have said that the conditions of Remark 2.1 or of Remark 2.2 are not only necessary but also sufficient for z_0 to be an extremal state.

3. In the case when Z is a finite-dimensional space the necessary tests for the extremal state may be somewhat sharpened.

Let z be a point of the set $\Omega \subset Z$ and let $l \in Z$. We say that the direction l is free of the points of the set Ω if there exist a neighborhood U of the element l and a number $\epsilon > 0$ such that from the relations $u \in U$, $0 < t \leq \epsilon$ it follows that $z + tu \notin \Omega$. The set $AC_\Omega(z)$ of all nonfree directions is, as is easily verified, a closed cone which is called the asymptotic cone of the set Ω at the point z .

The extremal state $z_0 \in \Omega$ of the system $[Z, \tau, \mathfrak{N}, \Omega]$ is called *strong* if the intersection $AC_\Omega(z_0) \cap M_{z_0} = \{0\}$.

It is easy to see that an extremal state is strong if there exists a cone $M' \subset Z$ such that every nonzero element $l \in M_{z_0}$ is an interior point of cone M' , and for M' the requirement for determining an extremal state is fulfilled if in it we replace M_{z_0} by M' . In particular, z_0 is automatically a strong extremal state if the cone $M_{z_0} \setminus \{0\}$ is open.¹

THEOREM 3. *Let Z be a finite-dimensional space and τ a normed topology. If z_0 is a strong extremal state of the system $[Z, \tau, \mathfrak{N}, \Omega]$ such that the cone $AC_\Omega(z_0)$ and the cone M_{z_0} are convex, and, moreover, if M_{z_0} does not reduce to the null element, then there exists a nontrivial linear functional f on Z such that $f(l) \geq 0$ for any $l \in M_{z_0}$, and for every number $\epsilon > 0$ we can select a neighborhood V of the point z_0 such that*

$$(1) \quad f(z) < f(z_0) + \epsilon \|z\|, \quad z \in V \cap \Omega.$$

Proof. Since the cones $AC_\Omega(z_0)$ and M_{z_0} have only one common element, there exists a hyperplane H differing from the sets indicated. Clearly, H is the set of zeros of a certain linear functional f which must be chosen so that $f(z) \geq 0$ ($z \in M_{z_0}$) and $f(z) \leq 0$ ($z \in AC_\Omega(z_0)$).

Taking $z_0 = 0$ for simplicity, let us assume that there does not exist a neighborhood V for which (1) is satisfied. This means that we can find a sequence $\{z_n\}$ of elements from Ω which converges to zero and, moreover, $f(z_n) \geq \epsilon \|z_n\|$, $n = 1, 2, \dots$. By setting $l_n = z_n / \|z_n\|$ we have $f(l_n) \geq \epsilon$. We can consider that the sequence $\{l_n\}$ converges to some element $l \in Z$. Then we have $l \in AC_\Omega(z_0)$, but meanwhile $f(l) \geq \epsilon$.

¹ It is clear that the definitions stated have a meaning for any vector space Z with an arbitrary topology, satisfying the assumptions of §2.

Remark 3.1. Only an insignificant change is necessary in the proof of the theorem if Z is a reflexive separable normed space and τ is the weak topology in Z . Here additional assumptions of the type of conditions (a)–(c) of Theorem 2 must be made with respect to the cones $AC_{\alpha}(z_0)$ and M_{z_0} .

Remark 3.2. In cases analogous to those specified in Remarks 2.1 and 2.2, we can supplement the results of the theorem by the assertion that $f(l) > 0$ for any $l \in M_{z_0}$ ($l \neq 0$).

Remark 3.3. By assuming as before that Z is finite-dimensional, we consider a point $z_0 \in \Omega$ such that the cone M_{z_0} is closed and there exists a linear functional f for which $f(l) > 0$ if $l \in M_{z_0}$ ($l \neq 0$) and which satisfies (1). As is not difficult to show, z_0 is a strong extremal state in this case.

4. Not infrequently, when the optimizable system $[Z, \tau, \mathfrak{M}, \Omega]$ is being considered, the set Ω or even the set Ω_0 is given as the image of some set X under a certain mapping $F: \Omega = F[X]$. In problems where such a situation is encountered the elements of set Ω are called the ordinary states of the system, while the elements of the set X are called controls. An element $x \in X$ is called an *extremal control* if $F(x)$ is an extremal state.

It is evident that since different controls may lead to one and the same state, every extremal state is determined, in general, by many extremal controls. On the basis of the nature of extremal states presented above we can make a number of conjectures on extremal controls.

5. Let us illustrate the concepts set forth by an example of a linear economic model.

The factors to be taken into account in the model (various kinds of raw materials, goods, finished products, industrial capacities, manufactured and natural resources as well as different collections of products and services which ensure the satisfaction of the specific needs of the public, of their associations, or of society on the whole) are called the *ingredients*. Here, thanks to the temporal and territorial aspects, one and the same factor can give rise to a whole series of different ingredients. The technoeconomic level of development of the system achieved is described in the model by listing the allowable *production methods*. Each of them is characterized by a vector whose components indicate the total productivity (or expenditure, in case the component has a negative sign) of the corresponding ingredients when a given method is used with unit intensity. The economic plan (the control) is determined by the selection of a vector with nonnegative components indicating the intensity of application of the different production methods. Because of the hypothesis of linearity, the results achieved here are characterized by a linear combination of vectors corresponding to the individual production methods, with coefficients equal to the intensities.

Suppose that n different ingredients participate in the model and that

the allowable production methods are described by the vectors $a^i = (a_{i1}, \dots, a_{in})$, $i = 1, 2, \dots, m$. Then, the controls are to be chosen from the set

$$(2) \quad X = \{x = (x_1, x_2, \dots, x_m) : x_i \geq 0, i = 1, 2, \dots, m\},$$

while the results achieved (the states) are characterized by the elements of the set

$$\Omega = \left\{ a(x) = \sum_{i=1}^m x_i a^i : x \in X \right\}.$$

As the cone of desirable directions at each point $a \in \Omega$ we can take one and the same cone,

$$(3) \quad M = \{l = (l_1, l_2, \dots, l_n) : l_j \geq 0, j \in J, \sum_{j \in J'} l_j > 0\},$$

where $J = \{1, 2, \dots, n\}$, while J' is some nonempty subset of this set.

THEOREM 4. *The state $a \in \Omega$ is extremal if and only if there exists a vector*

$$(4) \quad y = (y_1, y_2, \dots, y_n)$$

satisfying the conditions:

- (i) $y_j \geq 0, j \in J, y_j > 0, j \in J'$;
- (ii) $(a^i, y) \leq 0, i = 1, 2, \dots, m$;
- (iii) $(a, y) = 0$.

If here the state a corresponds to the control $x = (x_1, x_2, \dots, x_m) \in X$, condition (iii) can be rewritten in the language of control as

$$(iii') \quad (a^i, y) = 0 \text{ if } x_i > 0.$$

6. We go on to the analysis of convex economic models. In the case considered here, as also in §5, n ingredients participate in the model and the controls are to be chosen from set (2). However, the states achieved are now characterized by the elements of the set

$$\Omega = \{a(x) = (f_1(x), f_2(x), \dots, f_n(x)) : x \in X\},$$

where the f_j are fixed concave functions defined on the set X . As the cones of desirable directions here also we take one and the same cone (3).

THEOREM 5. *Let the point $a^0 \in \Omega$ be such that, for any nonzero vector (4) with components $y_j \geq 0, j \in J, \min_{j \in J'} y_j = 0$, there holds the inequality*

$$(a^0, y) < \sup_{a \in \Omega} (a, y).$$

In order that a^0 be an extremal state it is necessary and sufficient that there exist a vector (4) satisfying the conditions:

- (i) $y_j \geq 0, j \in J, y_j > 0, j \in J'$;
- (ii) $(a^0, y) = \max_{a \in \Omega} (a, y)$.

If here the state a^0 corresponds to the control $x^0 = (x_1^0, x_2^0, \dots, x_m^0) \in X$, and if, further, the functions f_j are continuously differentiable, then condition (ii) can be replaced by the following:

$$(ii') \quad \sum_{j \in J} \frac{\partial f(x^0)}{\partial x_i} y_j \leq 0, \quad i = 1, 2, \dots, m;$$

moreover, equality is attained in these inequalities only for those indices i for which $x_i^0 > 0$.

7. Let us now study the simplest continuous model of production and transport. Consider the metric bicomactum K with metric $r(x, y)$. Let $p_j, j = 1, 2, \dots, m$, denote expenditures of various kinds associated with the manufacture of a single product at the point $x \in K$, and let $p_j(x, y), j = m + 1, m + 2, \dots, n$, denote the expenditures of corresponding kinds associated with the transport of the single product from point x to point y . We shall assume that the functions introduced satisfy the relations:

$$(5) \quad p_j(y) - p_j(x) \leq c_j r(x, y), \quad 0 \leq p_j(x, y) \leq c_j r(x, y), \quad x, y \in K,$$

where the $c_j, j = 1, 2, \dots, n$, are fixed positive numbers.

The production and transport plan (the control) is determined by the selection of a nonnegative countably additive function ϕ given on the system B of all Borel sets of bicomactum K , and by the selection of a nonnegative function ψ , countably additive in each argument, given on $B \times B$. The results (states) achieved here are characterized by the countably additive function α and by the vector $a = (a_1, a_2, \dots, a_n)$ such that

$$\begin{aligned} \alpha(e) &= \phi(e) + \psi(K, e) - \psi(e, K), \quad e \in B, \\ a_j &= \begin{cases} \int_K p_j(x) \phi(de), & j = 1, 2, \dots, m, \\ \int_K \int_K p_j(x, y) \psi(de, de'), & j = m + 1, m + 2, \dots, n. \end{cases} \end{aligned}$$

As above, we denote the set of all states by Ω . As the cone of desirable directions at each point $\langle \alpha, a \rangle \in \Omega$ we take one and the same cone $M = \{\langle \beta, l \rangle\}$, where the β are nonnegative countably additive functions, and the $l = (l_1, l_2, \dots, l_n)$ are nonzero vectors with nonnegative components.

THEOREM 6. *The controls determined by the functions ϕ and ψ are extremal if and only if there exist a function u defined on K and a vector $v = (v_1, v_2, \dots, v_n)$ with positive components such that*

- (i) $0 \leq u(x) \leq \sum_{j=1}^m p_j(x) v_j, \quad x \in K,$
- (ii) $u(y) - u(x) \leq \sum_{j=m+1}^n p_j(x, y) v_j, \quad x, y \in K,$
- (iii) $u(x) = \sum_{j=1}^m p_j(x) v_j$ if $\phi(e) > 0$ for any neighborhood e of point x ,

(iv) $u(y) - u(x) = \sum_{j=m+1}^n p_j(x, y)v_j$ if $\psi(e, e') > 0$ for any neighborhoods e and e' of the points x and y , respectively.

8. Questions relating to extremal states are encountered also in economic models of consumption.

In such cases the space Z is interpreted as the space of commodities, so that an element $z \in Z$ is to be understood as the collection of commodities offered to the consumer. The set Ω is interpreted as the budget set, namely, the totality of those commodity collections each of which the consumer is in a position to acquire. Finally, the family \mathfrak{M} characterizes the consumer's tastes in the sense that if the collection z differs sufficiently little from collection z_0 , and here $z - z_0 \in M_{z_0}$, then the consumer prefers to acquire the collection z instead of z_0 (if, of course, it is allowed by his budgetary means).

It is natural to consider that the actual purchase by the consumer is that commodity collection $z_0 \in \Omega$ which is the extremal state of the system $[Z, \tau, \mathfrak{M}, \Omega]$ (here, τ is any topology in Z), and conversely, every extremal state z_0 can be realized (if the consumer is shown the collection z_0 , he will buy it).

Under the assumptions made the hypothesis of Theorem 1 can be explained as the known stability of the consumer's tastes: the direction of his interests do not depend too strongly on the commodity collection offered.

In the case when the hypothesis of Theorem 2 or of Theorem 3 is fulfilled (this fact is typically true for problems with real meaning), the functional f which occurs in these theorems plays the role of the consumption cost.

9. In conclusion we illustrate the total concept of §1-§3 by an example of the classical problem of a conditional extremum.

Let the $m + 1$ continuously differentiable functions ϕ, g_1, \dots, g_m be given in an open set Ω_0 of the Euclidean space R^n . Let us denote $\Omega = \{z \in \Omega_0: g_k(z) = 0, k = 1, 2, \dots, m\}$. It is required to find the point $z_0 \in \Omega$ at which the function ϕ has a (local) maximum.

Let us take a point $z \in \Omega_0$ and set

$$M_z^0 = \{u \in R^n: (\text{grad } \phi(z), u) > 0\}, \quad M_z = M_z^0 \cup \{0\},$$

$$\mathfrak{M} = \{M_z\}, \quad z \in \Omega_0.$$

We assume that $z_0 \in \Omega$ is the point giving the solution of the stated extremal problem. It is clear that z_0 will be an extremal state of the system $[R^n, \mathfrak{M}, \Omega]$ and, by virtue of what was said in §4, will even be strong. If $\text{grad } g_k(z_0) \neq 0$, $k = 1, \dots, m$, then the asymptotic cone of the set Ω at the point z_0 will be the intersection $\bigcap_{k=1}^m H_k$ of the hyperplanes

$$H_k = \{u \in R^n: (\text{grad } g_k(z_0), u) = 0\}.$$

Therefore, if $\text{grad } \phi(z_0) \neq 0$, then by Theorem 3, all of whose conditions are

obviously fulfilled, we can find a functional f having the properties indicated in the theorem. From Remark 3.2 it is clear that $f = \lambda \operatorname{grad} \phi(z_0)$, where $\lambda > 0$. Since for $u \in AC_\Omega(z_0)$, we should have $f(u) \leq 0$ and, consequently, since $f(u) = 0$ in view of the linearity of the set $AC_\Omega(z_0)$, we can say that the relations $(\operatorname{grad} g_k(z_0), u) = 0$, $k = 1, \dots, m$, imply the equality $(\operatorname{grad} \phi(z_0), u) = 0$ which, as is well known, is possible only if $\operatorname{grad} \phi(z_0)$ is a linear combination of the $\operatorname{grad} g_k(z_0)$, $k = 1, 2, \dots, m$, i.e., if there exist numbers $\lambda_1, \lambda_2, \dots, \lambda_m$ such that

$$(6) \quad \operatorname{grad} \phi(z_0) = \sum_{k=1}^m \lambda_k \operatorname{grad} g_k(z_0).$$

The latter equality holds even if $\operatorname{grad} \phi(z_0) = 0$ (when $\lambda_1 = \lambda_2 = \dots = \lambda_m = 0$). If we introduce the function $F = \phi - \sum_{k=1}^m \lambda_k g_k$, then by writing (6) as $\operatorname{grad} F(z_0) = 0$, we obtain the well-known necessary condition for a local relative extremum.²

² For example, see V. I. Smirnov, *Boundary value problems, integral equations and partial differential equations*, A Course of Higher Mathematics, vol. IV, Addison-Wesley, Reading, Massachusetts, 1964.

CLASSICAL SOLUTIONS OF DIFFERENTIAL EQUATIONS WITH MULTIVALUED RIGHT-HAND SIDE*

A. F. FILIPPOV†

Abstract. Several existence theorems are proved for absolutely continuous solutions and for continuously differentiable solutions of the equation $dx(t)/dt \in F(t, x(t))$, where $F(t, x)$ is a set continuously dependent on t, x .

1. A relation of the form

$$(1) \quad \frac{dx}{dt} \in F(t, x)$$

is called a differential equation with multivalued right-hand side, where $x = x(t)$ is an unknown n -dimensional vector-valued function, while $F(t, x)$ is a given multivalued function or, more precisely, a function which associates with every point (t, x) from a certain region of an $(n + 1)$ -dimensional space a set $F(t, x)$ of an n -dimensional space. The following, in particular, lead to relations of form (1): differential equations of the form $f(t, x, \dot{x}) = 0$ ($\dot{x} = dx/dt$); differential inequalities $f(t, x, \dot{x}) \geq 0$ (or $|\dot{x} - f(t, x)| \leq \epsilon$) [1]; contingent equations [2]; automatic control systems described by equations of the form

$$(2) \quad \dot{x} = f(t, x, u), \quad u \in Q,$$

where $x = x(t)$ and $u = u(t)$ are the unknown functions and for each t the value of the function $u(t)$ should belong to a given set Q (see [3]–[5]).

In this article we shall employ the following definitions for the solution of (1): (i) a classical solution is the vector-valued function $x(t)$ having a continuous derivative and satisfying (1) everywhere; (ii) a solution is the absolutely-continuous vector-valued function $x(t)$ satisfying (1) almost everywhere.

We shall prove theorems for the existence of solutions and of classical solutions without assumptions on the convexity and boundedness of the set $F(t, x)$ and on the possibility of picking out a continuous single-valued branch of the multivalued function $F(t, x)$. We shall study questions on the approximation of any solution by classical solutions and on the disposition of classical solutions in a funnel formed by solutions issuing from one point.

* Originally published in *Vestnik Moskovskogo Universiteta*, no. 3 (1967), pp. 16–26. Submitted on January 14, 1966. This translation into English has been prepared by N. H. Choksy.

Translated and printed for this Journal under a grant-in-aid by the National Science Foundation.

† Chair of Differential Equations, Moscow University, Moscow, USSR.

2. In what follows, sets will be denoted by capital letters, numbers and points in the n -dimensional space, by lower-case letters; $|x| = (x_1^2 + \dots + x_n^2)^{1/2}$; O is the origin; $\dot{x} = dx/dt$; μ is the Lebesgue measure on the real line; a.e. stands for almost everywhere. Further, ρ is the distance between points or sets; $\alpha(F, G)$ is the Hausdorff deviation between sets F and G , i.e.,

$$\alpha(F, G) = \max \left\{ \sup_{x \in F} \rho(x, G), \sup_{y \in G} \rho(y, F) \right\};$$

$\text{conv } F$ is the convex closure of the set F , i.e., the smallest convex closed set containing F .

In the following we shall usually assume that the function $F(t, x)$ satisfies the following conditions:

- (a) $F(t, x)$ is a nonempty closed set;
 - (b) the function $F(t, x)$ is continuous in t, x , i.e.,
- $$(3) \quad \alpha(F(t', x'), F(t, x)) \rightarrow 0 \quad \text{as } t' \rightarrow t, \quad x' \rightarrow x;$$
- (c) there exists a summable function $k(t)$ such that for any t, x, x' ,
- $$(4) \quad \alpha(F(t, x'), F(t, x)) \leq k(t) |x' - x|.$$

3. It is well known [2], [4] that if conditions (a) and (b) are fulfilled and if the set $F(t, x)$ is convex for any t, x , then a solution of (1) with initial condition $x(t_0) = x_0$ exists, while if only conditions (a) and (b) are fulfilled, then quasitrajectories exist [6]. In the following we do not assume convexity for the set $F(t, x)$.

LEMMA 1. For any point x and any sets F and G ,

$$(5) \quad |\rho(x, F) - \rho(x, G)| \leq \alpha(F, G).$$

LEMMA 2. If the function $F(p)$ is continuous, then the function $\rho(p) = \rho(x, F(p))$ also is continuous.

This follows from Lemma 1 and from the definition of continuity for the function $F(p)$.

LEMMA 3. Let M be a closed set, let $Q(t)$ be a nonempty bounded closed set in n -dimensional space for $t \in M$, and let the function $Q(t)$ be continuous (or upper semicontinuous with respect to inclusion [5]). Then a single-valued measurable function $u(t)$ exists such that $u(t) \in Q(t)$ for every $t \in M$.

This is a particular case, when $f(t) \equiv 0$ and $y(t) \equiv 0$, of the lemma in [5].

LEMMA 4. Let $F(t)$ satisfy conditions (a) and (b) and let the vector-valued function $w(t)$ be measurable. Then a measurable vector-valued function $u(t)$ exists such that for almost all t , $u(t)$ is a point in the set $F(t)$ nearest to $w(t)$, or one of the nearest if there are more than one of them.

Since $w(t)$ is measurable on the set M , $M = N + M_1 + M_2 + \dots$, $\mu N = 0$, all the M_i are closed sets, and on each M_i the function $w(t)$ is continuous over this set. For $t \in M_i$ let $Q(t)$ be the set of points of $F(t)$ nearest to the point $w(t)$. It is easy to see that the set $Q(t)$ is closed and bounded and that on M_i the function $Q(t)$ is upper semicontinuous with respect to inclusion. By virtue of Lemma 3, on each M_i we can construct a single-valued measurable function $u(t)$ such that $u(t) \in Q(t)$. The function $u(t)$ is measurable also on M .

THEOREM 1. *Let $F(t, x)$ satisfy conditions (a)–(c) in the region $t \in I$, $|x - y(t)| \leq b$, where I is a segment of the t -axis, the function $y(t)$ is absolutely continuous. Let $t_0 \in I$, let the function $\rho(t)$ be summable, and let*

$$(6) \quad |y(t_0) - x_0| \leq \delta < b, \quad \rho(\dot{y}(t), F(t, y(t))) \leq \rho(t) \quad \text{a.e.}$$

Then a solution $x(t)$ of the problem

$$(7) \quad \dot{x} \in F(t, x), \quad x(t_0) = x_0,$$

exists such that

$$(8) \quad |x(t) - y(t)| \leq \xi(t), \quad |\dot{x}(t) - \dot{y}(t)| \leq k(t)\xi(t) + \rho(t) \quad \text{a.e.},$$

$$(9) \quad \xi(t) = \delta e^{m(t)} + \left| \int_{t_0}^t e^{m(t)-m(s)} \rho(s) ds \right|, \quad m(t) = \left| \int_{t_0}^t k(r) dr \right|$$

(for $t \in I$ such that $\xi(t) \leq b$).

Proof. We construct the sequence of $x_i(t)$, $i = 1, 2, \dots$:

$$(10) \quad x_0(t) \equiv y(t), \quad x_{i+1}(t) = x_0 + \int_{t_0}^t v_i(s) ds, \quad i = 0, 1, 2, \dots,$$

where $v_i(t)$ is a measurable vector-valued function such that, for almost all t ,

$$(11) \quad v_i(t) \in F(t, x(t)), \quad |v_i(t) - \dot{x}_i(t)| = \rho(\dot{x}_i(t), F(t, x_i(t))).$$

The function $v_i(t)$ exists by virtue of Lemma 4. It is summable since $\dot{x}_i = v_{i-1} \in F(t, x_{i-1}(t))$, and by virtue of (4) the right-hand side of (11) does not exceed $k(t)|x_{i-1} - x_i|$ (or $\rho(t)$ when $i = 1$). Therefore,

$$(12) \quad \dot{x}_{i+1} = v_i, \quad |\dot{x}_{i+1}(t) - \dot{x}_i(t)| \leq k(t), \quad |x_i(t) - x_{i-1}(t)| \quad \text{a.e.},$$

$i = 1, 2, \dots$

It follows from (10), (11) and (6) that

$$(13) \quad |\dot{x}_1(t) - \dot{x}_0(t)| \leq \rho(t) \quad \text{a.e.},$$

$$|x_1(t) - x_0(t)| \leq \delta + \left| \int_{t_0}^t \rho(s) ds \right|.$$

From (12) and (13) by induction we obtain

$$(14) \quad \begin{aligned} & | \dot{x}_{i+1}(t) - \dot{x}_i(t) | \\ & \leq k(t) \left\{ \delta \frac{[m(t)]^{i-1}}{(i-1)!} + \left| \int_{t_0}^t \frac{[m(t) - m(s)]^{i-1}}{(i-1)!} \rho(s) ds \right| \right\} \quad \text{a.e.,} \end{aligned}$$

$$(15) \quad \left| x_{i+1}(t) - x_i(t) \right| \leq \delta \frac{[m(t)]^i}{i!} + \left| \int_{t_0}^t \frac{[m(t) - m(s)]^i}{i!} \rho(s) ds \right|$$

(in order to obtain (15) from (14) we need to compare the derivative of the right-hand side of (15) with (14) and to consider that both sides of (15) equal zero when $t = t_0$).

Adding together inequalities (15) for $i = 0, 1, \dots, j-1$, and using the inequality $1 + z/1! + \dots + z^j/j! \leq e^z, z \geq 0$, and (9), we obtain

$$(16) \quad |x_j(t) - y(t)| \leq \xi(t).$$

Hence, all the $x_i(t)$ have been determined for $t \in I$ such that $\xi(t) \leq b$.

By virtue of (14) and (15) the sequences of $x_i(t)$ and $\dot{x}_i(t) \equiv v_{i-1}(t)$ converge to the functions $x(t)$ and $v(t)$. Passing to the limit as $i \rightarrow \infty$ in (10) and (11) and taking (3) into account, we get that $x(t)$ is absolutely continuous and serves as the solution of problem (7).

COROLLARY. *If $F(t, x)$ satisfies conditions (a)–(c) in the region $t \in I, |x - x_0| \leq b$, then the solution $x(t)$ of problem (7) exists on the segment I or on a part of it.*

Proof. The hypothesis of Theorem 1 will be fulfilled if we take $y(t) \equiv x_0$; the function $\rho(t) = \rho(O, F(t, x_0))$ is continuous by Lemma 2.

THEOREM 2. *Let $x_0(t)$ be a solution of problem (7); in the region $R: t_0 \leq t \leq t_0 + a, |x - x_0(t)| \leq \epsilon_0$, the function $F(t, x)$ satisfies conditions (a), (b), (c) with $k(t) \equiv k = \text{const}$. Then for any $\epsilon > 0$ the solution $x(t)$ of problem (7) exists, having a bounded derivative $\dot{x}(t)$ and satisfying the inequality $|x(t) - x_0(t)| < \epsilon$ on the segment $I: t_0 \leq t \leq t_0 + a$.*

Proof. We take any $\epsilon, 0 < \epsilon < \epsilon_0$. By virtue of Lemma 2, $\max_R \rho(O, F(t, x)) = r < \infty$. We can find a number $b \geq k\epsilon, b \geq r$, such that the integral of $|\dot{x}(t)|$ over the set B of those points of segment I , where $|\dot{x}(t)| > b$, does not exceed $\epsilon/2e^{ka}$. Let

$$(17) \quad v(t) = \begin{cases} \dot{x}_0(t) & \text{for } t \in I - B, \\ 0 & \text{for } t \in B, \end{cases} \quad y(t) = x_0 + \int_{t_0}^t v(s) ds.$$

Then

$$(18) \quad |y(t) - x_0(t)| \leq \frac{\epsilon}{2e^{ka}}, \quad \rho(y(t), F(t, y(t))) = \rho(t) \quad \text{a.e.,}$$

$$(19) \quad \rho(t) \leq r \leq b \quad \text{for } t \in B; \quad \rho(t) \leq \frac{k\epsilon}{2e^{ka}} \quad \text{for } t \in I - B;$$

the estimate when $t \in I - B$ is obtained with the aid of the relation $\dot{y}(t) = \dot{x}_0(t) \in F(t, x_0(t))$, of inequality (4) and of the first formula in (18).

A solution of problem (7) satisfying inequality (8) exists by Theorem 1. By virtue of the choice of the set B and of formulas (9) and (19), we have

$$\int_B b \, ds < \int_B |\dot{x}_0(s)| \, ds \leq \frac{\epsilon}{2e^{ka}},$$

$$\xi(t) < \int_B e^{ka} r \, ds + \int_{t_0}^t e^{kt-ks} \frac{k\epsilon \, ds}{2e^{ka}} < \epsilon - \frac{\epsilon}{2e^{ka}}.$$

Now, with the help of (8) and (18) we obtain $|x(t) - x_0(t)| < \epsilon$. From (19) and the second inequality in (8) we have almost everywhere

$$|\dot{x} - \dot{y}| < k\epsilon \leq b \quad \text{for } t \in I - B,$$

$$|\dot{x} - \dot{y}| < k\epsilon + b \leq 2b \quad \text{for } t \in B.$$

Taking (17) into account we obtain $|\dot{x}(t)| < 2b$ for almost all $t \in I$.

4. Along with the equation $\dot{x} \in F(t, x)$ let us consider the equation $\dot{x} \in K(t, x)$, where $K(t, x) = \text{conv } F(t, x)$. Since $F(t, x) \subseteq K(t, x)$ always, every solution (or classical solution) of the equation $\dot{x} \in F(t, x)$ will be a solution (classical solution) of the equation $\dot{x} \in K(t, x)$. It is clear that the converse is not true. In [6] it was proven that under conditions (a), (b) and the boundedness of set $F(t, x)$, the limit of any convergent sequence of solutions of the equation $\dot{x} \in F(t, x)$ is a solution of the equation $\dot{x} \in K(t, x)$ and that every solution of the equation $\dot{x} \in K(t, x)$ is the limit of a sequence of approximate solutions of the equation $\dot{x} \in F(t, x)$.

Let us show that if condition (c) also is fulfilled the approximate solutions can be replaced by solutions.

THEOREM 3. *Let a function $x_0(t)$ be given and in the region R (see Theorem 2) let the function $F(t, x)$ satisfy conditions (a)–(c), and let the set $F(t, x)$ be bounded for all $(t, x) \in R$. If $x_0(t)$ is a solution of the equation $\dot{x} \in K(t, x)$, where $K(t, x) = \text{conv } F(t, x)$, then there exists a sequence of solutions of equation $\dot{x} \in F(t, x)$ which converges uniformly to $x_0(t)$ on the segment $I: t_0 \leq t \leq t_0 + a$.*

Proof. By virtue of [2, Theorem 1] there exists an m such that $|v| \leq m$ for all $v \in F(t, x)$ for all $(t, x) \in R$. According to [6] there exists a sequence of absolutely continuous functions $y_n(t)$ for which $y_n(t) \rightarrow x_0(t)$ uniformly on I as $n \rightarrow \infty$,

$$|\dot{y}_n(t)| \leq m, \quad \rho(\dot{y}_n(t), F(t, y_n(t))) = \rho_n(t) \rightarrow 0 \quad \text{a.e.}$$

Since $|\rho_n(t)| = |\dot{y}_n - v|$, $v \in F(t, x)$, we have $|\rho_n(t)| \leq 2m$. By virtue of Theorem 1, for all sufficiently large n there exists a solution $x_n(t)$ of (1)

for which

$$x_n(t_0) = x_0(t_0), \quad |x_n(t) - y_n(t)| \leq \xi_n(t);$$

$\xi_n(t)$ is determined from (9) with $\delta = |y_n(t_0) - x_0(t_0)|$, $\rho(t) = \rho_n(t)$; hence $\xi_n(t) \rightarrow 0$, $x_n(t) \rightarrow x_0(t)$ uniformly on I .

A similar theorem was proven in [7] for systems of form (2).

5. Thus, a solution of problem (7) exists if $F(t, x)$ satisfies conditions (a)–(c) of §2. A classical solution may or may not exist, as in the following example.

For every t , $0 < |t| \leq 1$, and for all $x = (x_1, x_2)$, let the set $F(t, x)$ be an arc of an ellipse in the (v_1, v_2) -plane,

$$v_1 = \cos \phi, \quad v_2 = t \sin \phi, \quad t^{-1} \leq \phi \leq t^{-1} + 2\pi - |t|,$$

while for $t = 0$ the set $F(t, x)$ is the segment $|v_1| \leq 1, v_2 = 0$. Then, the function $F(t, x)$ is independent of x and satisfies conditions (a)–(c) of §2. Consequently, for any $x_0 = (x_{01}, x_{02})$ the equation $\dot{x} \in F(t, x)$ has a solution which satisfies the initial condition $x(-1) = x_0$.

However, for any $t_1 > 0$ not even one classical solution of this equation exists in the interval $[-t_1, t_1]$ because there does not exist a vector-valued function $v(t)$ for which $v(t) \in F(t, x)$ for all $t \in [-t_1, t_1]$.

We state sufficient conditions for the existence of a classical solution of problem (7), which are different from those stated in [1].

LEMMA 5. *Let v be a point and let F and G be nonempty closed convex sets, f and g are the points of F and G closest to the point v , $\rho(v, F) = d$, $\alpha(F, G) = h$. Then $\rho(f, g) \leq \sqrt{h^2 + 4hd}$.*

Proof. If $v \in F$, then $f = v$, $d = 0$, $\rho(f, g) = \rho(v, G) \leq h$. If $v \notin F$, then the convex set F lies to one side of the plane P drawn through the point f perpendicular to the segment fv . Hence, the set G lies to the same side of the plane Q obtained by shifting P through a distance h from f to v . Furthermore, since $\alpha(F, G) = h$, in the set G we can find a point $g_1 \in S(f, h) \subseteq S(v, d + h)$. Here $S(a, r)$ is the closed sphere with center a and radius r . Consequently, $g \in S(v, d + h)$ also. Hence, g is contained in the segment intercepted on $S(v, d + h)$ by the plane Q . Since the radius of the segment equals $d + h$, the height is $2h$ and the point f is located in the middle of the segment's height, the distance from f to the furthest points of the segment is $\sqrt{h^2 + 4hd}$.

LEMMA 6. *Let $F(t, x)$ be a nonempty closed convex set, continuously dependent on t, x (see (3)); v_0 is some fixed point; $f(t, x)$ is the point of the set $F(t, x)$ closest to point v_0 . Then the function $f(t, x)$ is continuous.*

The assertion ensues from Lemma 5.

THEOREM 4. *If $F(t, x)$ satisfies the hypothesis of Lemma 6 in the neighbor-*

hood of the point (t_0, x_0) , then a classical solution of the problem

$$(20) \quad \dot{x} \in F(t, x), \quad x(t_0) = x_0, \quad \dot{x}(t_0) = v_0,$$

exists for any $v_0 \in F(t_0, x_0)$.

For example, such a solution is the solution of the Cauchy problem $\dot{x} = f(t, x)$, $x(t_0) = x_0$, where the function $f(t, x)$ is the same one as in Lemma 6.

THEOREM 5. *Let the function $F(t, x)$ satisfy condition (a) in a neighborhood U of the point (t_0, x_0) and let a constant k exist such that for any two points $(t, x) \in U$, $(t', x') \in U$,*

$$(21) \quad \alpha(F(t', x'), F(t, x)) \leq k |t' - t| + k |x' - x|.$$

Then a classical solution of problem (20) exists for any $v_0 \in F(t_0, x_0)$.

Proof. We take an $a > 0$ such that $ak < 1$ and such that the cone

$$t_0 \leq t \leq t_0 + a, \quad |x - x_0| \leq m(t - t_0), \quad \text{where} \quad m = \frac{|v_0| + ak}{1 - ak},$$

is contained in U . For every $j = 1, 2, \dots$ we construct the Euler polygons on the interval $J[t_0, t_0 + a]$. Let

$$h = h_j = 2^{-j}a, \quad t_i = t_0 + ih, \quad i \leq 2^j; \quad x_j(t_0) = x_0.$$

We construct $x_j(t)$ successively on the intervals $[t_0, t_1]$, $[t_1, t_2]$, \dots thus:

$$x_j(t) = x_j(t_i) + (t - t_i)v_{ji}, \quad t_i \leq t \leq t_{i+1}, \quad i = 0, 1, 2, \dots,$$

where $v_{j0} = v_0$, while when $i \geq 1$, v_{ji} is a point of the set $F(t_i, x_j(t_i))$ closest to the point $v_{j,i-1}$ (or any one of the closest ones if there are several). By induction we can prove that, when $t_i \leq t \leq t_{i+1}$, $i = 0, 1, \dots, 2^j - 1$,

$$(22) \quad |x_j(t) - x_j(t_i)| \leq m(t - t_i) \leq mh, \quad |x_j(t) - x_0| \leq m(t - t_0),$$

$$(23) \quad |v_{j,i+1} - v_{ji}| \leq k(h + mh),$$

$$|v_{j,i+1}| \leq |v_0| + (i + 1)hk(m + 1) \leq m.$$

Hence, the graphs of the functions $x_j(t)$ are contained in the cone we have constructed. Let $v_j(t)$ be a function equal to v_{ji} when $t = t_i$, $i = 0, 1, \dots, 2^j$, and linear in the intervals $[t_{i-1}, t_i]$. By virtue of (23), the functions $v_j(t)$, $j = 1, 2, \dots$, are uniformly bounded and equicontinuous on J . From them let us select a uniformly convergent subsequence $v_{j_n}(t) \rightarrow v(t)$. Since, for $t \in [t_i, t_{i+1}]$,

$$\dot{x}_j(t) = v_{ji} = v_j(t_i), \quad |\dot{x}_j(t) - v_j(t)| \leq |v_{j,i+1} - v_{ji}| \leq k(h + mh)$$

($\dot{x}_j(t_i)$ is a right derivative, $\dot{x}_j(t_{i+1})$, a left), when $j = j_n$, $n \rightarrow \infty$,

$$\dot{x}_j(t) \rightarrow v(t), \quad x_j(t) \rightarrow x(t) \equiv x_0 + \int_{t_0}^t v(s) ds;$$

the convergence is uniform on J . Hence, the function $\dot{x}(t) \equiv v(t)$ is continuous, $x(t_0) = x_0$, $\dot{x}(t_0) = v_0$.

If when $j = j_n$, $n \rightarrow \infty$, $t = \text{const.}$, we pass to the limit in the relation

$$\dot{x}_j(t) = v_j(t_i) = v_{ji} \in F(t_i, x_j(t_i)), \quad i = i(j, t), \quad t_i \leq t \leq t_{i+1},$$

and use inequalities (21), (22), we obtain $\dot{x}(t) \in F(t, x(t))$.

We shall say that the set $F(t, x)$, depending on t, x , is *uniformly locally connected* if there exists a function $\eta(\lambda)$, $0 < \lambda < \infty$, such that $\eta(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$ and that for any t, x , any two points of the set $F(t, x)$, the distance between which is less than λ , can be joined by a connected set contained within the set $F(T, x)$ and having a diameter less than $\eta(\lambda)$. We note that in this case the set $F(t, x)$ is connected.

THEOREM 6. *Let $x_0(t)$ be a solution of problem (7), the function $F(t, x)$ in the region R (see Theorem 2) satisfy conditions (a), (b), (c) with $k(t) \equiv k = \text{const.}$, and the set $F(t, x)$ be uniformly locally connected. Then for any $\epsilon > 0$ and $v_0 \in F(t_0, x_0)$ there exists a classical solution of problem (20) on the segment $I[t_0, t_0 + a]$ satisfying the inequality $|x(t) - x_0(t)| < \epsilon$.*

Proof. We take any ϵ , $0 < \epsilon < \epsilon_0$, and with the aid of Theorem 2 we construct a solution $x_1(t)$ of problem (7) such that

$$(24) \quad |x_1(t) - x_0(t)| < \frac{\epsilon}{2}, \quad |\dot{x}_1(t)| < 2b \quad \text{a.e.}, \quad 2b \geq |v_0|.$$

We take $d > \eta(8b)$, where the function η is the same one as above, and we also take λ_i , ϵ_i , μ_i such that

$$(25) \quad 0 < \lambda_i < \lambda_{i-1}, \quad 2\lambda_1 = 8b < d, \quad 2\lambda_i \leq \eta(2\lambda_i) < 2^{1-i}d, \\ i = 2, 3, 4, \dots,$$

$$(26) \quad 0 < \epsilon_i < \frac{\epsilon}{2^i(1 + 3ae^{ka})}, \quad \epsilon_i < \frac{\lambda_{i+1}}{1 + 6e^{ka}}, \quad 0 < \mu_i < \frac{2^{i-2}\epsilon_i}{(k+1)d}.$$

We construct the sequence of solutions $x_i(t)$ of problem (20) such that

$$(27) \quad |x_i(t) - x_0(t)| < (1 - 2^{-i}\epsilon), \\ |\dot{x}_i(t)| < 2b + 4d - 2^{3-i}d \quad \text{a.e.},$$

and such that the discontinuity in the function $\dot{x}_i(t)$ is less than λ_i , i.e.,

$$(28) \quad \limsup_{t' \rightarrow t} \dot{x}_i(t') - \liminf_{t' \rightarrow t} \dot{x}_i(t') < \lambda_i, \quad t_0 \leq t \leq t_0 + a.$$

Let us assume that the solution $x_i(t)$ with these properties has been constructed and let us construct the solution $x_{i+1}(t)$. We take $\delta_i > 0$ such

that for any $t_1, t_2, |t_1 - t_2| < \delta_i, |z_1 - z_2| < (2b + 4d)\delta_i$ we have

$$(29) \quad \alpha(F(t_1, z_1), F(t_2, z_2)) < \epsilon_i,$$

$$(30) \quad |\dot{x}_i(t_1) - \dot{x}_i(t_2)| < \lambda_i + \epsilon_i \quad \text{a.e.}$$

Then when $|t_1 - t_2| < \delta_i$, by virtue of (27) and (29) we have

$$(31) \quad \alpha(F(t_1, x_i(t_1)), F(t_2, x_i(t_2))) < \epsilon_i.$$

Since $\dot{x}_i(t)$ is measurable and $\dot{x}_i(t) \in F(t, x_i(t))$ almost everywhere on I , we can find a closed set $M \subseteq I$ such that $t_0 \notin M$, $\mu(I - M) < \mu_i$, the length of each interval of which $I - M$ is composed is less than δ_i , the function $\dot{x}_i(t)$ is continuous on the set M over this set, and $\dot{x}_i(t) \in F(t, x_i(t))$ everywhere on M .

We construct the function $v(t)$. Let $v(t_0) = v_0$, $v(t) = \dot{x}_i(t)$ for $t \in M$, $v(t)$ is linear on the intervals $[\beta, \gamma]$, which are such that $\beta, \gamma \in M$, $(\beta, \gamma) \subseteq I - M$, $|v(\beta) - v(\gamma)| < \epsilon_i$. Then $v(t)$ is continuous on the set on which it is yet to be defined and which is composed of the segment I minus a finite number of intervals. Let (β, γ) be any one of them. As $v(\gamma - 0)$ we take the point of the set $F_1 = F(\beta, x_i(\beta))$ closest to the point $v(\gamma) = \dot{x}_i(\gamma) \in F(\gamma, x_i(\gamma))$. Since $\gamma - \beta < \delta_i$, by virtue of (30) and (31),

$$(32) \quad |v(\gamma) - v(\beta)| < \lambda_i + \epsilon_i, \quad |v(\gamma - 0) - v(\gamma)| < \epsilon_i.$$

Making use of the uniform local connectedness, we join the points $v(\beta) \in F_1$ and $v(\gamma - 0) \in F_1$ by the connected set $C \subseteq F_1$ with diameter $d_i < \eta(\lambda_i + 2\epsilon_i)$ (see (32)). By virtue of (25), (26),

$$(33) \quad 2\epsilon_i < \lambda_{i+1} < \lambda_i, \quad d_i < \eta(2\lambda_i) < 2^{1-i} d.$$

Let us construct the points $w_j \in C, j = 1, 2, \dots, s$, such that

$$w_1 = v(\beta), \quad w_s = v(\gamma - 0), \quad |w_j - w_{j+1}| < \epsilon_i.$$

We set $v(t) = w_j$ when $t_{j-1} \leq t < t_j, j = 1, 2, \dots, s$, where $t_j = \beta + j(\gamma - \beta)/s$. We define $v(t)$ analogously on all such intervals (β, γ) . If there is an interval $(\beta^*, \gamma^*) \subseteq I - M$ in which $\gamma^* = t_0 + a$, then we set $v(t) = v(\beta^*)$ on $[\beta^*, \gamma^*]$. Now $v(t)$ has been defined on I and has only a finite number of discontinuities, the jumps at the points of discontinuity being less than ϵ_i .

Let us estimate $v(t) - \dot{x}_i(t)$. On M , $v = \dot{x}_i$; on those intervals (β, γ) where $v(t)$ is linear, $|v(t) - v(\beta)| < \epsilon_i < 2^{-i} d$. On the remaining intervals $(\beta, \gamma), |v(t) - v(\beta)| < 2^{1-i} d$ by virtue of (33). If $\beta \neq t_0$, then $\beta \in M$; if $\beta = t_0, i \geq 2$, then $v(\beta) = v_0 = \dot{x}_i(\beta)$. Since $0 < \gamma - \beta < \delta_i$,

by virtue of (30) when $t \in (\beta, \gamma)$,

$$(34) \quad v(\beta) = \dot{x}_i(\beta) \in F(\beta, x_i(\beta)), \quad |\dot{x}_i(\beta) - \dot{x}_i(t)| < \lambda_i + \epsilon_i \quad \text{a.e.}$$

The last inequality is true also when $i = 1$ by virtue of (24) and (25). From the relations written down it follows that

$$(35) \quad \begin{aligned} |v(t) - \dot{x}_i(t)| &< 2^{1-i}d + \lambda_i + \epsilon_i \quad \text{for } t \in I - M, \\ v(t) &= \dot{x}_i(t) \quad \text{for } t \in M. \end{aligned}$$

Let $\rho^*(t) = \rho(v(t), F(t, x_i(t)))$. On the set M , $\rho^*(t) = 0$. On the intervals where the function $v(t)$ is linear, $|v(t) - v(\beta)| < \epsilon_i$ and, by virtue of (34),

$$(36) \quad \rho(v(t), F(\beta, x_i(\beta))) < \epsilon_i, \quad 0 < t - \beta < \gamma - \beta < \delta_i.$$

On the remaining intervals (β, γ) we have $v(t) \in C \subseteq F(\beta, x_i(\beta))$. Hence (36) is valid when $t \in I - M$ and, by virtue of (31) and (5),

$$(37) \quad \begin{aligned} \rho^*(t) &< 2\epsilon_i \quad \text{for } t \in I - M, \quad \mu(I - M) < \mu_i, \\ \rho^*(t) &= 0 \quad \text{for } t \in M. \end{aligned}$$

Let us construct the absolutely continuous function $y(t)$ for which

$$(38) \quad y(t_0) = x_0 = x_i(t_0), \quad \dot{y}(t) = v(t) \quad \text{a.e.}$$

Since $\mu(I - M) < \mu_i$, from (35), (25) and (26) we have

$$(39) \quad |y(t) - x_i(t)| < 2^{2-i}d\mu_i < \frac{\epsilon_i}{k+1}.$$

From inequalities (37), (5), (4) and (39) it follows that $\rho(t) \equiv \rho(\dot{y}(t), F(t, y(t))) < 3\epsilon_i$. By Theorem 1 the solution $x_{i+1}(t)$ of problem (7) exists, for which on I ,

$$(40) \quad |x_{i+1}(t) - \dot{y}(t)| < 3\epsilon_i a e^{ka}, \quad |\dot{x}_{i+1}(t) - \dot{y}(t)| < 3\epsilon_i e^{ka} \quad \text{a.e.}$$

Since as $t \rightarrow t_0$ we have $\dot{y}(t) \rightarrow v_0$ and $y(t) \rightarrow x_0$, $\rho(t) \rightarrow 0$ and, by virtue of (9), $\dot{x}_{i+1}(t_0) = v_0$ exists. From (39), (40), (26) and (35), (38), (25), it follows that

$$(41) \quad |x_{i+1}(t) - x_i(t)| < 3\epsilon_i a e^{ka} + \epsilon_i < 2^{-i-1}\epsilon,$$

$$(42) \quad |\dot{x}_{i+1}(t) - \dot{x}_i(t)| < 2^{1-i}d + \lambda_i + \epsilon_i + 3\epsilon_i e^{ka} < 2^{2-i}d \quad \text{a.e.}$$

Since the discontinuities of the function $\dot{y} \equiv v$ are less than ϵ_i , by virtue of (40) and (26) the discontinuities of the functions \dot{x}_{i+1} are less than $\epsilon_i + 6\epsilon_i e^{ka} < \lambda_{i+1}$.

Thus, if the solution $x_i(t)$ possesses properties (27) and (28), then $x_{i+1}(t)$ also possesses them (in view of the estimates obtained). According to (24),

the solution $x_1(t)$ possesses these properties and, therefore, all the solutions $x_i(t)$ possess these properties and satisfy inequalities (41) and (42).

By virtue of (41) and (42) the sequence of $x_i(t)$ converges uniformly on I , while that of $\dot{x}_i(t)$, almost everywhere on I . Passing to the limit in the relations

$$x_i(t) = x_0 + \int_{t_0}^t \dot{x}_i(s) ds, \quad \dot{x}_i(t) \in F(t, x_i(t)) \quad \text{a.e.},$$

with the help of (4) we obtain

$$(43) \quad x(t) = x_0 + \int_{t_0}^t \lim \dot{x}_i(s) ds, \quad \dot{x}(t) = \lim \dot{x}_i(t) \in F(t, x(t)) \quad \text{a.e.}$$

From (42), (43) and (30), for almost all t_1, t_2 , $|t_1 - t_2| < \delta_i$, we have

$$|\dot{x}(t_1) - \dot{x}_i(t_1)| < 2^{3-i}d, \quad |\dot{x}(t_1) - \dot{x}(t_2)| < 2^{4-i}d + \lambda_i + \epsilon_i.$$

Hence, $\dot{x}(t)$ is uniformly continuous on the set $I - N$, $\mu N = 0$. Hence from (43) it follows also that $\dot{x}(t)$ exists and is continuous for all $t \in I$.

Let us show that $\dot{x}(t_0) = v_0$. When $0 < t - t_0 < \delta^*$ we have

$$|\dot{x}(t_0) - \dot{x}(t)| < \epsilon^*, \quad |\dot{x}(t) - \dot{x}_i(t)| < 2^{3-i}d \quad \text{a.e.}$$

Hence from (38), (40) it follows that

$$(44) \quad |\dot{x}(t_0) - v(t)| < \epsilon^* + 2^{3-i}d + 3\epsilon_i e^{ka} \quad \text{a.e.}$$

Passing to the limit as $t \rightarrow t_0$ and noting that $v(t) \rightarrow v_0$, we get that $|\dot{x}(t_0) - v_0|$ is not larger than the right-hand side of (44), which can be made as small as desired. Hence, $\dot{x}(t_0) = v_0$.

6. For a given equation $\dot{x} \in F(t, x)$, the set H of points of the (t, x) -space lying on the solutions of problem (7), is called a funnel at the point (t_0, x_0) . Let us show that under certain conditions the classical solutions fill a set everywhere dense in H .

THEOREM 7. (i) *Let $F(t, x)$ satisfy the hypothesis of Theorem 6 in a region D containing every solution $x(t)$ of problem (7) together with a certain neighborhood $|x - x(t)| < \epsilon^*$ of it on the interval $I [t_0, t_1]$. Then, for any $v_0 \in F(t_0, x_0)$ and $t^* \in I$, the set $M(t^*)$ of points $x(t^*)$ lying on the classical solutions of problem (20) is everywhere dense in the section $H(t^*)$ of the funnel at the point (t_0, x_0) by the plane $t = t^*$.*

(ii) *Further, let the set $F(t, x)$ be bounded, $K(t, x) = \text{conv } F(t, x)$, and let the segment $t_0 \leq t \leq t_1$ of the funnel H_K at the point (t_0, x_0) for the equation $\dot{x} \in K(t, x)$, be contained inside D . Then the set $M(t^*)$ is everywhere dense also in the section $H_K(t^*)$ of the funnel H_K by the plane $t = t^*$.*

Proof. (i) Let $x^* \in H(t^*)$. By the definition of a funnel, there exists a

solution $x_0(t)$ of problem (7) such that $x_0(t^*) = x^*$. For any $\epsilon > 0$, by Theorem 6 we can find the classical solution $x(t)$ of problem (20), for which $|x(t) - x_0(t)| < \epsilon$ when $t_0 \leq t \leq t^*$. Assertion (ii) follows from (i) because by virtue of Theorem 3 the set $H(t^*)$ is everywhere dense in $H_K(t^*)$.

Let us show that, in general, the set $M(t^*)$ does not fill up the whole section $H(t^*)$ of the funnel. For linear systems, considered in [3, Chap. 3], the section of the funnel¹ is convex and the time-optimal control always is on the boundary of the funnel section. By virtue of [3, §17, Theorem 9], all the optimal controls except for a finite number of them are discontinuous. Hence, the solutions with discontinuous derivatives corresponding to the optimal controls (and only these by virtue of the uniqueness theorem [3, §18]) hit all the points (except a finite number) on the boundary of the funnel section. Let us show that the classical solutions fill the whole interior of the funnel.

THEOREM 8. *Let $A(t)$ and $B(t)$ be matrices, $U(t)$ be a closed convex set, A, B, U depend continuously on t . Then the classical solutions of the system*

$$(45) \quad \dot{x} = A(t)x + B(t)u \quad \text{for } u \in U(t), \quad x(t_0) = x_0,$$

fill the whole interior of the funnel.

Proof. Any point x^* lying inside $H(t^*)$ can be surrounded by points $z_i \in H(t^*)$, $i = 1, \dots, m$, such that x^* will be inside a polyhedron with vertices at z_i . By Theorem 7 we can find the classical solutions $x_i(t)$, $u_i(t)$ (with continuous $\dot{x}_i(t)$) of problem (45), for which $|x_i(t^*) - z_i|$ are so small that x^* lies inside a polyhedron with vertices at $x_i(t)$. Then

$$x^* = c_1 x_1(t^*) + \dots + c_m x_m(t^*), \quad c_i \geq 0, \quad c_1 + \dots + c_m = 1.$$

Obviously, $x(t) = \sum c_i x_i(t)$, $u(t) = \sum c_i u_i(t)$ is the classical solution of problem (45), hitting on the point x^* when $t = t^*$.

For nonlinear systems the classical solutions may not fall at certain interior points of the funnel section. Indeed, let not even one classical solution hit on a certain segment of the boundary of section $H(t)$ of the funnel. As t increases, the section $H(t)$ is deformed continuously. This boundary segment can jut into the interior, for example, such that the boundary turns into a circle and one of the radii. Then at the points of the radius, which are interior points of the funnel section, no classical solutions hit.

THEOREM 9. *In system (2) let $x = (x^1, \dots, x^n)$, let f and $\partial f / \partial x^i$ be continuous and, for any t, x , let the function f map the set Q onto a closed, strictly*

¹ For system (2) the name funnel is given to the set of points (t, x) which lie on trajectories with the initial condition $x(t_0) = x_0$, corresponding to all possible admissible controls.

convex set $R(t, x)$ (see [5]). Then the classical solutions fill the whole funnel.

Proof. We take any point (t_1, x_1) of the funnel and consider the problem: find the solution $x(t)$ of system (2), satisfying conditions $x(t_0) = x_0$, $x(t_1) = x_1$, such that the integral of $|\dot{x}(t)|^2$ over the interval $t_0 \leq t \leq t_1$ has the least possible value. This integral is a lower semicontinuous functional of $x(t)$; hence, a solution exists [8]. Since $|\dot{x}(t)|^2 = |f(t, x, u)|^2$, the solution satisfies the maximum principle [3, §8]. In the notation of [3], $f = (f^1, \dots, f^n)$, $|f|^2 = (f^1)^2 + \dots + (f^n)^2 = f^0$, and for each t the control $u(t)$ should be such that the scalar product

$$(46) \quad \psi_0 f^0(t, x, u) + \dots + \psi_n f^n(t, x, u)$$

has the largest possible value for $u \in U = Q$. When u ranges over U , f ranges over $R(t, x)$, while (f^0, f^1, \dots, f^n) ranges over a piece P of the paraboloid

$$f^0 = (f^1)^2 + \dots + (f^n)^2, \quad (f^1, \dots, f^n) \in R(t, x),$$

in the space of f^0, f^1, \dots, f^n . The maximum of the scalar product (46) is attained at the point f^* of the piece P , whose projection onto the vector $\psi = (\psi_0, \dots, \psi_n)$ takes its extreme position (in the direction of this vector). Since $\psi_0 \leq 0$ by virtue of Theorem 6 of [3], the vector ψ is directed toward the side of decrease of f^0 , i.e., toward the side of strict convexity of the piece P . Then, the point f^* will depend continuously on ψ even if it is found on the boundary of the piece P . Since $\psi(t) \neq 0$ and it depends continuously on t (see [3]), f^* and $\dot{x} = (f^1, \dots, f^n)$ will depend continuously on t .

REFERENCES

- [1] A. D. MYSHKIS, *Differential inequalities with locally bounded derivatives*, Uchen' Zap. Khar'k. Un-ta, 138; Zap. Mekh.-Mat. Fak. i Khar'k. Matem. O-va, 30 (1964), pp. 152-163.
- [2] E. A. BARBASHIN AND YU. I. ALIMOV, *Theory of relay differential equations*, Izv. Vyssh. Uchebn. Zaved. Matematika, no. 1 (1962), pp. 3-13.
- [3] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.
- [4] T. WAŻEWSKI, *Systèmes de commande et équations au contingent*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 9 (1961), pp. 151-155.
- [5] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76-84.
- [6] T. WAŻEWSKI, *Sur une généralisation de la notion des solutions d'une équation au contingent*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 10 (1962), pp. 11-15.
- [7] A. TUROWICZ, *Sur les trajectoires et quasitrajectoires des systèmes de commande non-linéaires*, Ibid., 10 (1962), pp. 529-531.
- [8] A. F. FILIPPOV, *Differential equations with many-valued discontinuous right-hand side*, Soviet Math. Dokl., 4 (1963), pp. 941-945.

REMARKS ON SOME RECENT EXTENSIONS OF FILIPPOV'S IMPLICIT FUNCTIONS LEMMA*

MARC Q. JACOBS†

1. Introduction. The purpose of this note is to point out how implicit functions theorems obtained in [12] and [10] can be extended to infinite-dimensional situations. Implicit functions theorems are a fundamental part of proving what have come to be called "closure theorems" [5]. The implicit functions theorems which we state (Theorems 2.2 and 2.3) provide immediate extensions of Falb's closure theorem [8] which are suited for application to distributed parameter control systems. Theorem 2.1 is an extension of a result of Scorza Dragoni's [6], relating properties C_μ and C^* (see §2 for the terminology). G. Goodman [9] has called attention to Scorza Dragoni's result. Castaing [4] has recently given an extension of Scorza Dragoni's theorem in which the local compactness of the space X in Theorem 2.1 plays a fundamental role. We choose to utilize uniform continuity (or local uniform continuity) in our approach, and we obtain stronger results. Theorems 2.2 and 2.3 are extensions of results obtained by Castaing [3], [4] and McShane and Warfield [11], respectively.

2. Results. Throughout this note μ denotes a positive Radon measure [2, p. 41 ff.] defined on a compact Hausdorff space T , and (X, ρ_1) , (Y, ρ_2) denote metric spaces. We assume without loss of generality that the metric ρ_2 is bounded. A mapping $f: T \times X \rightarrow Y$ is given. We say that the mapping f has *property C^** if and only if for every $\epsilon > 0$ there is an open set $E_\epsilon \subset T$ such that $\mu(E_\epsilon) < \epsilon$ and such that $f| (T \setminus E_\epsilon) \times X$ is continuous. The mapping f is said to have *property C_μ* if and only if the mappings $f(t, \cdot): X \rightarrow Y$, $t \in T$, are continuous, and the mappings $f(\cdot, x): T \rightarrow Y$, $x \in X$, are measurable.

LEMMA 2.1. *Let the mapping f have property C_μ , and let X be separable. Then the mapping $\omega: T \times [0, \infty) \rightarrow R$ defined by*

$$\omega(t, \delta) = \sup \{ \rho_2(f(t, x_1), f(t, x_2)) \mid \rho_1(x_1, x_2) \leq \delta \}, \quad \delta \geq 0,$$

has the following properties:

- (i) $\omega(\cdot, \delta)$ is measurable for each $\delta \geq 0$,
- (ii) $\omega(t, \rho_1(x_1, x_2)) \geq \rho_2(f(t, x_1), f(t, x_2))$ for $t \in T$, $x_1, x_2 \in X$.

* Received by the editors May 29, 1967, and in final revised form August 24, 1967.

† Department of Mathematics, Rice University, Houston, Texas 77001. This research was supported in part by the Air Force Office of Scientific Research, Office of Aerospace Research, U. S. Air Force, under AFOSR Grant 693-67, in part by the Army Research Office, Durham, under Contract DA-31-124-ARO-D-270 and in part by the National Science Foundation under Grant GP-5970.

Proof. For each $\delta > 0$ define J_δ to be the set

$$\{(x_1, x_2) \in X \times X \mid \rho_1(x_1, x_2) \leq \delta\}.$$

Since X is separable, $X \times X$ is separable, and consequently, J_δ is separable for every $\delta > 0$. Let $A_\delta = \{(\alpha_{n\delta}, \beta_{n\delta})\}$ denote a dense sequence in J_δ , $\delta > 0$. Define a mapping $\eta: T \times (0, \infty) \rightarrow R$ by the relation

$$\eta(t, \delta) = \sup_{\alpha_{n\delta}, \beta_{n\delta} \in A_\delta} \rho_2(f(t, \alpha_{n\delta}), f(t, \beta_{n\delta})), \quad t \in T, \quad \delta > 0.$$

Then, since f has property C_μ , and since ρ_2 is continuous, it follows that the mappings $t \in T \rightarrow \rho_2(f(t, \alpha_{n\delta}), f(t, \beta_{n\delta}))$ are measurable, $n = 1, 2, 3, \dots$, $\delta > 0$. Consequently, $\eta(\cdot, \delta)$ is measurable, $\delta > 0$. Evidently, $\eta(t, \delta) \leq \omega(t, \delta)$ for each $t \in T$ and $\delta > 0$. Suppose $\eta(t, \delta) < \omega(t, \delta)$. Then there exist $x_1^*, x_2^* \in X$ such that $\rho_1(x_1^*, x_2^*) \leq \delta$ and

$$(2.1) \quad \rho_2(f(t, x_1^*), f(t, x_2^*)) > \eta(t, \delta) \geq \rho_2(f(t, \alpha_{n\delta}), f(t, \beta_{n\delta})), \\ (\alpha_{n\delta}, \beta_{n\delta}) \in A_\delta.$$

There is, however, a sequence $\{(\alpha_{n\delta}^*, \beta_{n\delta}^*)\} \in A_\delta$ such that $\lim \alpha_{n\delta}^* = x_1^*$ and $\lim \beta_{n\delta}^* = x_2^*$. Therefore,

$$\lim \rho_2(f(t, \alpha_{n\delta}^*), f(t, \beta_{n\delta}^*)) = \rho_2(f(t, x_1^*), f(t, x_2^*)) \leq \eta(t, \delta),$$

which is contrary to (2.1). We conclude that $\eta(\cdot, \delta) = \omega(\cdot, \delta)$ for $\delta > 0$, thereby showing that $\omega(\cdot, \delta)$ is measurable for $\delta > 0$. Obviously the function $\omega(\cdot, 0) \equiv 0$ is measurable.

Conclusion (ii) is an immediate consequence of the definition of ω .

LEMMA 2.2. *Let the hypotheses of Lemma 2.1 remain in effect, and let $f(t, \cdot): X \rightarrow Y$ be uniformly continuous on X for each $t \in T$. Then the mappings $\omega(t, \cdot): [0, \infty) \rightarrow R$, $t \in T$, are continuous at 0.*

Proof. The conclusion follows directly from the uniform continuity of the mappings $f(t, \cdot)$, $t \in T$.

THEOREM 2.1 (Scorza Dragoni [6]). *Let the space X be separable, let the mapping f have property C_μ , and let the mappings $f(t, \cdot)$, $t \in T$, be uniformly continuous on X . Then f has property C^* .*

Proof. Let $Q = \{q_i\}$ be a sequence which is dense in X . Given $1 > \epsilon > 0$ and a positive integer i , there is an open set $E_{i\epsilon} \subset T$ such that $\mu(E_{i\epsilon}) < (\epsilon/2)^i/2$ and such that the mapping $f(\cdot, q_i): T \setminus E_{i\epsilon} \rightarrow Y$ is continuous. We note that $\mu(\bigcup_{i=1}^\infty E_{i\epsilon}) \leq \sum_{i=1}^\infty \mu(E_{i\epsilon}) < \epsilon/2$, that the set E_ϵ^* defined to be the set $\bigcup_{i=1}^\infty E_{i\epsilon}$ is open, and that $f(\cdot, q_i)$ is continuous on $T \setminus E_\epsilon^*$ for each positive integer i . Let the mapping $\omega: T \times [0, \infty) \rightarrow R$ be defined as in Lemma 2.1. Then by Lemma 2.1 the mappings $\omega(\cdot, \delta)$, $\delta \geq 0$, are measurable, and by Lemma 2.2 we have that $\omega(\cdot, 2/j) \rightarrow 0$ as $j \rightarrow \infty$. Consequently, by Egorov's theorem [2, p. 187] there is an open set $E_\epsilon^* \subset T$

such that $\mu(E_\epsilon^*) < \epsilon/2$ and such that $\omega(\cdot, 2/j) \rightarrow 0$ uniformly on $T \setminus E_\epsilon^*$ as $j \rightarrow \infty$. If E_ϵ is defined to be the set $E_\epsilon^* \cup E_\epsilon^*$, then E_ϵ is open and $\mu(E_\epsilon) < \epsilon$. We shall prove that $f|(T \setminus E_\epsilon) \times X$ is continuous. Let $(t', x') \in (T \setminus E_\epsilon) \times X$, and $h > 0$ be given. Since $f(t', \cdot)$ is continuous at x' , it follows that there is an $r_0 > 0$ such that $\rho_1(x, x') < r_0$ implies

$$(2.2) \quad \rho_2(f(t', x'), f(t', x)) < h/3.$$

Choose a positive integer j_0 such that $2/j_0 < r_0$, and choose a positive integer $J_0 \geq j_0$ such that $j \geq J_0$ implies that $0 \leq \omega(t, 2/j) < h/3$, for each $t \in T \setminus E_\epsilon$. There is a $q_{i_0} \in Q$ such that $\rho_1(x', q_{i_0}) < 1/J_0$. The functions $\omega(t, \cdot)$, $t \in T$, are nondecreasing. Hence we have that whenever $\rho_1(x, x') < 1/J_0$, then $\rho_1(x, q_{i_0}) \leq \rho_1(x, x') + \rho_1(x', q_{i_0}) < 2/J_0$, and consequently,

$$(2.3) \quad h/3 > \omega(t, 2/J_0) \geq \omega(t, \rho_1(x, q_{i_0})) \geq \rho_2(f(t, x), f(t, q_{i_0})), \quad t \in T \setminus E_\epsilon.$$

Since $f(\cdot, q_{i_0})$ is continuous on $T \setminus E_\epsilon$, it follows that there is a neighborhood $V(t')$ of t' such that $t \in (T \setminus E_\epsilon) \cap V(t')$ implies

$$(2.4) \quad \rho_2(f(t, q_{i_0}), f(t', q_{i_0})) < h/3.$$

If we combine (2.2), (2.3) and (2.4), then we have that $t \in V(t')$, $t \in T \setminus E_\epsilon$, and $\rho_1(x, x') < 1/J_0$ imply

$$\begin{aligned} \rho_2(f(t, x), f(t', x')) &\leq \rho_2(f(t', x'), f(t', q_{i_0})) + \rho_2(f(t', q_{i_0}), f(t, q_{i_0})) \\ &\quad + \rho_2(f(t, q_{i_0}), f(t, x)) < h/3 + h/3 + h/3 = h. \end{aligned}$$

We shall say that a mapping $h: X \rightarrow Y$ is *locally uniformly continuous* if and only if for each $x \in X$ there is a neighborhood of x (i.e., a set which contains an open set containing x) on which h is uniformly continuous. Of course, if h is continuous and X is locally compact, then h is locally uniformly continuous.

COROLLARY 2.1. *If f has property C_μ , if the mappings $f(t, \cdot)$, $t \in T$, are locally uniformly continuous, and if X is separable, then f has property C^* .*

COROLLARY 2.2. *If X is compact, and if f has property C_μ , then f has property C^* .*

COROLLARY 2.3. *If X is separable and locally compact, and if f has property C_μ , then f has property C^* .*

Corollaries 2.1 and 2.2 follow immediately from Theorem 2.1, and Corollary 2.3 follows from Corollary 2.1.

At this point it will be convenient to introduce some terminology concerning set-valued (or multivalued) mappings. We are referring to mappings F of a set S into the power set of another set A , i.e., into $2^A = \{M \mid M \subset A\}$. If A is a topological space, $C(A)$ denotes the collection of nonempty closed subsets of A , and $\mathbf{K}(A)$ denotes the collection of nonempty compact subsets of A . If F is a mapping, $F: S \rightarrow 2^A$, and $M \subset A$,

then $F \cap M$ will denote the mapping $s \in S \rightarrow F(s) \cap M$, $F^{-1}M$ will denote the set $\{s \in S \mid F(s) \cap M \neq \emptyset\}$, and $\mathfrak{D}(F)$ will denote the set $\{s \in S \mid F(s) \neq \emptyset\}$.

A mapping $F: T \rightarrow 2^X$ is *measurable* if and only if $F^{-1}M$ is measurable for every closed set $M \subset X$.

In the remainder of this note y is a measurable mapping, $y: T \rightarrow Y$, such that $y(t) \in f(t, X)$ for every $t \in T$, and $\Gamma: T \rightarrow 2^X \setminus \{\emptyset\}$ denotes the mapping defined by

$$\Gamma(t) = \{x \in X \mid f(t, x) = y(t)\}, \quad t \in T.$$

We observe that, if f has property C_μ , then $\Gamma(t) \in C(X)$ for every $t \in T$.

LEMMA 2.3. *If f has property C^* on $T \times K$, if K is a subset of X , then there is a set $N \subset T$ such that $\mu(N) = 0$, and there is a sequence T_n , $n = 1, 2, 3, \dots$, of compact subsets of T such that $\bigcup_{n=1}^\infty T_n = T \setminus N$ and such that $f \mid T_n \times K$ is continuous, $n = 1, 2, 3, \dots$.*

Proof. We can select open sets $E_n \subset T$ such that $\mu(E_n) < 1/n$ and such that $f \mid (T \setminus E_n) \times K$ is continuous, $n = 1, 2, 3, \dots$. The sets $N = \bigcap_{n=1}^\infty E_n$, $T_n = T \setminus E_n$, $n = 1, 2, 3, \dots$, fulfill the conditions of the lemma.

LEMMA 2.4. *If f has property C_μ , if the mappings $f(t, \cdot)$, $t \in T$, are locally uniformly continuous, and if X is a Souslin space [1, p. 124], then Γ is measurable.*

Proof. Since X is separable [1, Prop. 4, p. 125], Corollary 2.1 applies to give that f has property C^* . By Lemma 2.3 there is a sequence T_n of compact subsets of T , and a set of measure zero, $N \subset T$, such that $\bigcup_{n=1}^\infty T_n = T \setminus N$ and such that $f \mid T_n \times X$ is continuous, $n = 1, 2, 3, \dots$. Thus $\Gamma \mid T_n$ is measurable, $n = 1, 2, 3, \dots$ [3, p. 410], and consequently if M is a closed subset of X , then

$$\Gamma^{-1}M = [N \cap \Gamma^{-1}M] \cup \left[\bigcup_{n=1}^\infty (\Gamma \mid T_n)^{-1}M \right]$$

is measurable.

THEOREM 2.2. *If f has property C_μ , if the mappings $f(t, \cdot)$, $t \in T$, are locally uniformly continuous, and if X is a Polish space [1, p. 121], then there is a measurable function $x: T \rightarrow X$ such that $x(t) \in \Gamma(t)$ for each $t \in T$.*

Proof. This is an immediate consequence of Lemma 2.4 and Theorem 3 (see [3]).

LEMMA 2.5. *If $K \in \mathbf{K}(X)$, and if f has property C_μ , then $\Gamma \cap K$ is measurable.*

Proof. The mapping $f: T \times K \rightarrow Y$ has property C^* by Corollary 2.2. By Lemma 2.3 there is a sequence T_n of compact subsets of T , and a set $N \subset T$ of measure zero, such that $\bigcup_{n=1}^\infty T_n = T \setminus N$ and such that $f \mid T_n \times K$ is continuous, $n = 1, 2, 3, \dots$. If M is a closed subset of X , then

$$(\Gamma \cap K)^{-1}M = \{t \in N \mid \Gamma(t) \cap K \cap M \neq \emptyset\}$$

$$\cup \left[\bigcup_{n=1}^{\infty} \{t \in T_n \mid \Gamma(t) \cap K \cap M \neq \emptyset\} \right],$$

and the sets $\{t \in T_n \mid \Gamma(t) \cap K \cap M \neq \emptyset\}$, $n = 1, 2, 3, \dots$, are closed. Consequently, $(\Gamma \cap K)^{-1}M$ is measurable.

LEMMA 2.6. *If $K \in \mathbf{K}(X)$, and if f has property C_μ , then $\mathfrak{D}(\Gamma \cap K)$ is measurable. Moreover, if $\mathfrak{D}(\Gamma \cap K) \neq \emptyset$, then there is a measurable function $x_K: \mathfrak{D}(\Gamma \cap K) \rightarrow K$ such that $x_K(t) \in \Gamma(t)$ for every $t \in \mathfrak{D}(\Gamma \cap K)$.*

Proof. $\mathfrak{D}(\Gamma \cap K)$ is measurable by Lemma 2.5. If $\mathfrak{D}(\Gamma \cap K) \neq \emptyset$, then the existence of the measurable function x_K can be deduced from Theorem 2 (see [3]).

LEMMA 2.7. *If $Q = \bigcup_{n=1}^{\infty} K_n$, $K_n \in \mathbf{K}(X)$, $n = 1, 2, 3, \dots$, then $\mathfrak{D}(\Gamma \cap Q)$ is measurable. Moreover, if $\mathfrak{D}(\Gamma \cap Q) \neq \emptyset$, then there is a measurable function $x_Q: \mathfrak{D}(\Gamma \cap Q) \rightarrow Q$ such that $x_Q(t) \in \Gamma(t) \cap Q$ for every $t \in \mathfrak{D}(\Gamma \cap Q)$.*

Proof. Since $\mathfrak{D}(\Gamma \cap Q) = \bigcup_{n=1}^{\infty} \mathfrak{D}(\Gamma \cap K_n)$, it follows from Lemma 2.6 that $\mathfrak{D}(\Gamma \cap Q)$ is measurable. If $\mathfrak{D}(\Gamma \cap Q) \neq \emptyset$, then without loss of generality we may assume $\mathfrak{D}(\Gamma \cap K_n) \neq \emptyset$, $n = 1, 2, 3, \dots$, and again by Lemma 2.6 there is a sequence of measurable functions $x_n: \mathfrak{D}(\Gamma \cap K_n) \rightarrow K_n$, such that, for each n , $x_n(t) \in \Gamma(t) \cap K_n$ for each $t \in \mathfrak{D}(\Gamma \cap K_n)$. The sets $\mathfrak{D}(\Gamma \cap K_n) \setminus \bigcup_{i=1}^{n-1} \mathfrak{D}(\Gamma \cap K_i)$, $n = 1, 2, 3, \dots$, form a measurable partition of $\mathfrak{D}(\Gamma \cap Q)$. We define a mapping $x_Q: \mathfrak{D}(\Gamma \cap Q) \rightarrow Q$ by the rule that $x_Q(t) = x_n(t)$ if $t \in \mathfrak{D}(\Gamma \cap K_n) \setminus \bigcup_{i=1}^{n-1} \mathfrak{D}(\Gamma \cap K_i)$. The mapping x_Q is measurable and has the desired properties.

We shall say that a function $x: T \rightarrow X$ is *quasi-measurable* if and only if $x^{-1}(K)$ is measurable for every $K \in \mathbf{K}(X)$. x is measurable implies x is quasi-measurable [2, p. 191]. If X is the union of a countable number of compact sets, then the two concepts are equivalent.

If the continuum hypothesis is invoked, then the following theorem is obtained.

THEOREM 2.3 (McShane-Warfield [11]). *Let X be separable, and let f have property C_μ . Then there is a quasi-measurable function $x: T \rightarrow X$ such that $x(t) \in \Gamma(t)$ for every $t \in T$.*

Proof. Let c denote the cardinality of the continuum. By the continuum hypothesis the first uncountable ordinal Ω is preceded by c ordinals. The cardinality of $C(X)$ is c [11, p. 46]. Consequently, there is an ordinal number $\Omega' \leq \Omega$ such that $\mathbf{K}(X) = \{K_\alpha \mid \alpha \text{ an ordinal number, } \alpha < \Omega'\}$. For each $\alpha < \Omega'$ we define $Q_\alpha = \bigcup_{\beta \leq \alpha} K_\beta$. Then Q_α is the union of countably many compact sets. For each $\alpha < \Omega'$ define sets T_α by

$$T_\alpha = (\Gamma^{-1}Q_\alpha) \setminus (\Gamma^{-1} \bigcup_{\beta < \alpha} Q_\beta).$$

By Lemma 2.7 the sets T_α , $\alpha < \Omega'$, are measurable. The collection $\{T_\alpha \mid \alpha < \Omega'\}$ forms a measurable partition of T . By Lemma 2.7 there are measurable mappings $x_\alpha: T_\alpha \rightarrow Q_\alpha$ such that $x_\alpha(t) \in \Gamma(t)$ for every $t \in T_\alpha$. A mapping $x: T \rightarrow X$ is defined by the condition that $x(t) = x_\alpha(t)$, if $t \in T_\alpha$. The proof that x is quasi-measurable follows precisely as in [11].

REFERENCES

- [1] N. BOURBAKI, *Topologie générale*, Hermann, Paris, 1958, Chap. IX.
- [2] ———, *Intégration*, Hermann, Paris, 1952, Chap. I-IV.
- [3] C. CASTAING, *Quelques problèmes de mesurabilité liés à la théorie de la commande*, C. R. Acad. Sci. Paris, 262 (1966), pp. 409-411.
- [4] ———, *Sur les multi-applications mesurables*, Rev. Française d'Informatique et de Recherche Operationnelle, 1 (1967), pp. 91-126.
- [5] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints. I and II*, Trans. Amer. Math. Soc., 124 (1966), pp. 369-412, 413-429.
- [6] G. SCORZA DRAGONI, *Un teorema sulle funzioni continue rispetto ad una e misurabili rispetto ad un'altra variabile*, Rend. Sem. Mat. Univ. Padova, 17 (1948), pp. 102-106.
- [7] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1957.
- [8] P. FALB, *Infinite dimensional control problems 1: on the closure of the set of attainable sets for linear systems*, J. Math. Anal. Appl., 9 (1964), pp. 12-22.
- [9] G. GOODMAN, *An application of quasi-continuity in the sense of Scorza Dragoni to an existence question in the theory of optimal control*, Proc. Conference on the Mathematical Theory of Control, 1967, at The University of Southern California, Los Angeles, Academic Press, New York, to be published.
- [10] M. Q. JACOBS, *Attainable sets in linear systems with unbounded controls*, Proc. Conference on the Mathematical Theory of Control, 1967, at The University of Southern California, Los Angeles, Academic Press, New York, to be published.
- [11] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41-47.
- [12] C. OLECH, *A note concerning set-valued measurable functions*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 317-321.

FUNCTIONS OF RELAXED CONTROLS*

J. WARGA†

1. Introduction. The mathematical control theory deals primarily with functionals defined in terms of ordinary differential or difference equations. It is our purpose here, and in a paper to follow, to extend certain methods and results of the control theory to a more general setting. In particular, we wish to generalize certain results of [1] and [2]. In this endeavor, and especially in arguments pertaining to problems of existence, we extend certain concepts first introduced by L. C. Young in his study of “generalized curves” [3], [4].

Let \mathcal{R} be a class of mappings from a set T to a set R , and let x be a mapping from \mathcal{R} into some topological space H . In typical problems of the mathematical control theory, T is a closed interval of the real axis, R and H are Euclidean sets, and $x(\rho)$ is the endpoint of a curve defined by a system of ordinary differential equations involving the control function ρ . Our present investigations are, however, motivated by more general problems that may involve mappings x defined, for example, by partial differential equations or by nonadditive set functions.

In studying variational problems involving ordinary differential equations, L. C. Young [3] has introduced the very fruitful concept of a “generalized” curve. This concept basically involves replacing a time dependent vector in Euclidean n -space E_n by a time dependent “averaging” operator acting on continuous functions. Thus a derivative $\dot{\xi}$ of a function ξ from an interval $[t_0, t_1]$ to E_n is viewed by Young as a linear functional that maps, for each $t \in T = [t_0, t_1]$, a continuous function c on E_n into the number $\int_{t_0}^t c(\xi(\tau)) d\tau$. We have later [1] followed an analogous approach in studying certain aspects (proper representations) of “standard” control problems, and in studying minimax problems [5].

In the present paper we somewhat modify this approach. With T and R assumed to be metric and compact, and an appropriate positive measure defined on T , we consider a measurable function ρ from T to R as a fixed linear functional (independent of t) whose domain is the space \mathcal{B} of functions ϕ on $T \times R$, continuous on R for every $t \in T$, measurable on T for every $r \in R$, and with $\sup_{r \in R} |\phi(t, r)|$ integrable on T . Viewed as an element of the dual space \mathcal{B}^* of \mathcal{B} , ρ maps ϕ into the number $\int_T \phi(t, \rho(t))$. This

* Received by the editors May 22, 1967, and in final revised form August 28, 1967.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This research was supported by the National Aeronautics and Space Administration under Grant NGR 22-011-020.

modified definition is partly motivated by the fact that T need no longer be one-dimensional (or finite-dimensional), and partly by the desire to simplify the use of our results in certain applications.

If we define \mathfrak{R} to be the class of all measurable mappings from T to R , we can imbed it in a larger set \mathfrak{S} of "measurable relaxed controls". We define \mathfrak{S} as the class of regular Borel probability measures on R , and \mathfrak{S} as the set of measurable functions from T to \mathfrak{S} . Again, we identify an element σ of \mathfrak{S} with the bounded linear functional in \mathfrak{B}^* that maps ϕ into $\int_T \int_R \phi(t, r) \sigma(dr; t)$. (Here $\sigma(R'; t)$ designates the $\sigma(t)$ -measure of a subset R' of R .) In this sense, \mathfrak{S} is a subset of \mathfrak{B}^* , and we prove that \mathfrak{S} is the closure of \mathfrak{R} in the weak star topology of \mathfrak{B}^* , and that it is sequentially compact.

In many problems we wish to restrict the mappings ρ to the set \mathfrak{R}^* of measurable functions from T to R with the property that $\rho(t) \in R^*(t)$ on T . Here R^* is an appropriately defined "measurable" mapping from T to the class of nonempty subsets of R . We then define \mathfrak{S}^* to be the subset of \mathfrak{S} with elements σ such that the measure $\sigma(t)$ is supported on $\bar{R}^*(t)$ (the closure of $R^*(t)$) for all $t \in T$. We prove that \mathfrak{S}^* is the closure of \mathfrak{R}^* and that it is sequentially compact in the weak star topology of \mathfrak{B}^* . Thus, whenever the function x , defined on \mathfrak{R}^* , can be continuously extended to \mathfrak{S}^* , we are assured of the existence of "relaxed" solutions in typical optimization problems, and we can approximate these relaxed solutions with ordinary controls (in \mathfrak{R}^*). As an illustration, we consider in §3 and §5 the "standard" problem defined by ordinary differential equations.

The next task is the derivation of necessary conditions for minimum. We shall discuss that subject, for both relaxed and ordinary controls, in a paper to follow.

2. Existence and approximation theorems. We shall henceforth assume that T and R are compact metric spaces, that B_0 is a compact topological space, and B_1 is a closed set in a topological space H . We shall assume, further, that a nonnegative, finite, regular, complete, and nonatomic measure is defined on T . We represent by ρ , and sometimes by $\rho(\cdot)$, a mapping from T to R , and by $\rho(t)$ the image of a point t under the mapping. A similar distinction is consistently made between a function (mapping) and the image of a particular point under the mapping. The symbol $|r_1, r_2|$ designates the distance of points r_1 and r_2 in R , and similarly $|t', t|$ will designate the distance in T . A mapping ρ is "measurable" if, for every $\epsilon > 0$, there exists a closed set F_ϵ in T , whose measure $|F_\epsilon|$ is at least $|T| - \epsilon$, and such that the restriction of ρ to F_ϵ is continuous.

The usual definition of continuity and the above definition of measurability can be easily seen to be equivalent to the following statement: ρ is continuous at t (on a set T_1), respectively measurable on T_1 , if the func-

tion ψ , defined by $\psi(t) = c(\rho(t))$ on T , is continuous at t (on T_1), respectively measurable on T_1 , for every choice of a continuous function c from R to E_1 (the real line).

Let R^* be a mapping from T to the class of nonempty subsets of R , let \mathfrak{R} be the space of measurable functions ρ from T to R , and let E_n be the Euclidean n -space. We are given a function $x^0: \mathfrak{R} \times B_0 \rightarrow E_1$ and a function $x: \mathfrak{R} \times B_0 \rightarrow H$. We wish to investigate the *original* problem of determining the minimum of $x^0(\rho, b)$, subject to the restrictions that $\rho(t) \in R^*(t)$ a.e. in T and $x(\rho, b) \in B_1$. Alternately, we wish to consider "approximate minimizing solutions" to this problem. An "approximate solution" α is a sequence $\{\rho_j, b_j\}_{j=1}^\infty$ such that the ρ_j are measurable functions from T to R , $b_j \in B_0$, $\rho_j(t) \in R^*(t)$ a.e. in T , $x^0(\rho_j, b_j)$, respectively $x(\rho_j, b_j)$, converges, as $j \rightarrow \infty$, to a number ξ_α^0 , respectively a point ξ_α , and $\xi_\alpha \in B_1$. An "approximate minimizing solution" is an approximate solution that minimizes ξ_α^0 .

Let now S be the class of regular probability measures defined on the Borel subsets of R . We shall refer to a function ρ from T to R as an "original control", and to a function σ from T to S as a "relaxed control". A relaxed control σ is "continuous", respectively "measurable", on a set T_1 if

$\int_R c(r) \sigma(dr; \cdot)$ is a continuous, respectively measurable, function on T_1

for every choice of a continuous function $c: R \rightarrow E_1$. Here $\sigma(R'; t)$ represents the $\sigma(t)$ -measure of a Borel set R' . We can easily verify that if σ is a measurable control and R' is a Borel subset of R , then $\sigma(R'; \cdot)$ is measurable. We shall denote by \mathfrak{S} the set of measurable relaxed controls.

If a relaxed control σ_ρ has the property that $\sigma_\rho(t)$ is a measure consisting of a single mass point $\rho(t)$ a.e. in T , then we refer to it, somewhat loosely but without any fear of confusion, as an original control ρ . In this sense we consider \mathfrak{R} to be a subset of \mathfrak{S} . We shall also treat original, respectively relaxed, controls as identical if they differ only on a set of measure 0 in T .

DEFINITION 2.1. We shall say that a function $y: \mathfrak{S} \times B_0 \rightarrow H$ is a *Young representation* of x if y coincides with x on $\mathfrak{R} \times B_0$. We similarly define $y^0: \mathfrak{S} \times B_0 \rightarrow E_1$ as a Young representation of x^0 .

Let now \mathfrak{B} be a Banach space of real-valued functions on $T \times R$ defined as follows: $\phi \in \mathfrak{B}$ if $\phi(\cdot, r)$ is measurable on T for every $r \in R$, $\phi(t, \cdot)$ is continuous on R for every $t \in T$, and there exists an integrable scalar function ϕ_{sup} on T such that $|\phi(t, r)| \leq \phi_{\text{sup}}(t)$ on $T \times R$. We define the norm in \mathfrak{B} by

$$\|\phi\| = \int_T \sup_{r \in R} |\phi(t, r)| \, dt.$$

We may clearly do so since $\sup_{r \in R} |\phi(t, r)|$ is measurable for every $\phi \in \mathfrak{B}$, as can be easily verified.

It is known [6, Theorem 11, p. 149 and Theorem 22, p. 117] that the above definition is equivalent to the statement that $\mathfrak{B} = L_T^1(C_R)$, where C_R is the Banach space of continuous real-valued functions c on R with the norm $|c| = \sup_{r \in R} |c(r)|$ and $L_T^1(C_R)$ is the Banach space of measurable functions f from T to C_R with the norm $|f| = \int_T |f(t)|_{C_R} < \infty$.

Let \mathfrak{B}^* be the dual space of \mathfrak{B} , and let $\langle l, \phi \rangle$ represent the value of $l \in \mathfrak{B}^*$ evaluated at $\phi \in \mathfrak{B}$. We can define \mathfrak{S} as a subset of \mathfrak{B}^* by setting

$$\langle \sigma, \phi \rangle = \int_T \int_R \phi(t, r) \sigma(dr; t).$$

This definition is permissible since it is known that $\int_R \phi(t, r) \sigma(dr; t)$ is measurable and integrable for every $\sigma \in \mathfrak{S}$ and $\phi \in \mathfrak{B}$.

Let L_T^∞ be the set of essentially bounded scalar measurable functions on T . We can define $L_T^\infty \times \mathfrak{S}$ as a subset of \mathfrak{B}^* by setting, for every $(f, \sigma) \in L_T^\infty \times \mathfrak{S}$,

$$\langle (f, \sigma), \phi \rangle = \int_T \int_R \phi(t, r) f(t) \sigma(dr; t).$$

DEFINITION 2.2. For the topology in \mathfrak{B}^* , we shall define convergence in \mathfrak{B}^* (hence also in \mathfrak{B} , \mathfrak{S} , $L_T^\infty \times \mathfrak{B}$, and $L_T^\infty \times \mathfrak{S}$) in terms of the weak star topology in \mathfrak{B}^* . Specifically, we shall say that a sequence l_1, l_2, \dots in \mathfrak{B}^* converges to l if

$$\lim_{j \rightarrow \infty} \langle l_j, \phi \rangle = \langle l, \phi \rangle \quad \text{for all } \phi \in \mathfrak{B}.$$

We next consider a mapping R^* from T into the collection of subsets of R satisfying the following assumption.

ASSUMPTION 2.3. For every $\epsilon > 0$ there exists a closed subset T_ϵ of T , of measure at least $|T| - \epsilon$, with the following properties:

(i) for every $\bar{t} \in T_\epsilon$ and every $r \in \bar{R}^*(\bar{t})$ (the closure of $R^*(\bar{t})$) there exists a measurable original control ρ , continuous at \bar{t} when restricted to T_ϵ , and such that $|\rho(\bar{t}), r| < \epsilon$ and $\rho(t) \in R^*(t)$ on T ;

(ii) the mapping R^* , when restricted to T_ϵ , is continuous with respect to the Hausdorff distance of sets, that is, for every \bar{t} in T_ϵ and every $h > 0$, there exists $\delta = \delta(h, \bar{t})$ such that $R^*(t) \subset U(R^*(\bar{t}), h)$ and $R^*(\bar{t}) \subset U(R^*(t), h)$ if $t \in T_\epsilon$ and $|t, \bar{t}| < \delta$. Here $U(R', h)$ is the open h -neighborhood of a set R' in R .

Let $\mathfrak{R}^* = \{\rho \in \mathfrak{R} \mid \rho(t) \in R^*(t) \text{ on } T\}$, $\mathfrak{S}^* = \{\sigma \in \mathfrak{S} \mid \sigma(\bar{R}^*(t); t) = 1 \text{ on } T\}$, and $\mathfrak{B}^{**} = \{l \in \mathfrak{B}^* \mid \langle l, \phi_1 \rangle = \langle l, \phi_2 \rangle \text{ provided } \phi_1(t, r) = \phi_2(t, r) \text{ for all } t \in T \text{ and all } r \in \bar{R}^*(t)\}$. We can now state our basic approximation and existence theorems that we shall prove in §4.

THEOREM 2.4. Let the mapping R^* satisfy condition (i) of Assumption

2.3. Then S^* and \mathcal{B}^{**} are the closures of, respectively, \mathcal{R}^* and $L_T^\infty \times \mathcal{R}^*$. In particular, S and \mathcal{B}^* are the closures of, respectively, \mathcal{R} and $L_T^\infty \times \mathcal{R}$.

THEOREM 2.5. Let the mapping R^* satisfy (ii) of Assumption 2.3. Then the set S^* is sequentially compact.

Let y^0 and y be Young representations of, respectively, x^0 and x , and assume that y^0 and y are continuous on $S^* \times B_0$ (with respect to the product topology on $S^* \times B_0$). Let \hat{X} be the image of $\mathcal{R}^* \times B_0$ in $E_1 \times H$ under the mapping (x^0, x) , and let \hat{Y} be similarly the image of $S^* \times B_0$ under the mapping (y^0, y) . Then \hat{Y} is the closure of \hat{X} in $E_1 \times H$ if the mapping R^* also satisfies (i) of Assumption 2.3.

We indicate in §4.1 a procedure for approximating $\bar{\sigma} \in S^*$ with a sequence $\bar{\rho}_1, \bar{\rho}_2, \dots$ in \mathcal{R}^* .

As a corollary of Theorems 2.4 and 2.5, we derive the following theorem.

THEOREM 2.6. Let the assumptions be the same as in Theorem 2.5, and let X and Y be the images in H of, respectively, $\mathcal{R}^* \times B_0$ under x and of $S^* \times B_0$ under y . Then either $Y \cap B_1$ is empty, or there exist $\bar{\sigma} \in S^*$ and $\bar{b} \in B_0$ that yield the minimum of $y^0(\sigma, b)$ subject to the condition that $y(\sigma, b) \in B_1$. If the construction described in §4.1 is used to approximate $\bar{\sigma}$ with the sequence $\{\bar{\rho}_j\}_{j=1}^\infty$ in \mathcal{R}^* , then the sequence $\{\bar{\rho}_j, \bar{b}\}_{j=1}^\infty$ is a minimizing approximate solution of the original problem.

We next state an analogue of a lemma of Filippov [7, p. 78]. (See also [12] for an extension of this lemma by McShane and Warfield. It has been pointed out by the referee that Theorem 2.7 below can be deduced from results of C. Castaing [13].)

THEOREM 2.7. Let Assumption 2.3 be satisfied, let $\phi = (\phi^1, \dots, \phi^n) \in \mathcal{B} \times \dots \times \mathcal{B}$ and let $\bar{\sigma}$ be a (not necessarily measurable) mapping from T to S such that $\bar{\sigma}(\bar{R}^*(t); t) = 1$ on T and the function ψ , defined by $\psi(t) = \int_R \phi(t, r) \bar{\sigma}(dr; t)$, is measurable on T . Then there exists some (measurable) $\sigma \in S^*$ such that $\psi(t) = \int_R \phi(t, r) \sigma(dr; t)$ on T .

3. Control problems defined by ordinary differential equations. As an illustration of the way the results of the preceding section can be applied, we consider the following much discussed problem [7], [8], [1]: let T be the closed interval $[t_0, t_1]$ of the real axis, A a closed subset of E_n , B_0 a compact subset of A , $g = (g^1, \dots, g^n): E_n \times T \times R \rightarrow E_n$, $G^*(v, t) = \{g(v, t, r) \mid r \in R^*(t)\}$ ($v \in E_n$, $t \in T$), $F^*(v, t)$ the convex closure of $G^*(v, t)$, and $|g| = (\sum_{i=1}^n (g^i)^2)^{1/2}$. We shall say that an absolutely continuous function $\xi: T \rightarrow E_n$ is an "original curve" if $d\xi(t)/dt = \xi(t) = g(\xi(t), t, \rho(t))$ a.e. in T for some $\rho \in \mathcal{R}^*$, $\xi(t) \in A$ on T , and $\xi(t_0) \in B_0$. We shall say that ξ is a "relaxed curve" if $\xi(t) \in F^*(\xi(t), t)$ a.e. in T , $\xi(t) \in A$ on T , and $\xi(t_0) \in B_0$. Then the following theorem holds.

THEOREM 3.1. *Let the mapping R^* satisfy Assumption 2.3, let $g^i(v, \cdot, \cdot) \in \mathfrak{B}$ for $i = 1, \dots, n$ and for every $v \in E_n$, and assume that there exists an integrable scalar function ψ on T such that*

$$(3.1) \quad |g(v, t, r) - g(v', t, r)| \leq \psi(t) |v - v'|$$

for $v \in E_n$, $v' \in E_n$, $t \in T$ and $r \in R$. Then the set \tilde{Y} of relaxed curves is compact in the topology of uniform convergence and every relaxed curve can be uniformly approximated by curves ξ_j , $j = 1, 2, \dots$, such that $\dot{\xi}_j(t) = g(\xi_j(t), t, \rho_j(t))$ a.e. in T , where $\rho_j \in \mathfrak{R}^$, $j = 1, 2, \dots$, (i.e., $\rho_j(t) \in R^*(t)$ on T and ρ_j is measurable).*

Furthermore, every relaxed curve ξ corresponds to a relaxed control $\sigma \in \mathfrak{S}^$ such that $\dot{\xi}(t) = \int_R g(\xi(t), t, r) \sigma(dr; t)$ a.e. in T .*

This theorem generalizes in some respects the results of Filippov [7], Roxin [8], and Warga [1].

4. Proofs of the approximation and existence theorems.

DEFINITION 4.1. We shall say that P_τ is a *dense sequence of partitions* of T (see [9, pp. 171–174]) if $P_\tau = \{P_\tau^1, P_\tau^2, \dots\}$; $P_\tau^i = \{T_1^i, T_2^i, \dots, T_{j_i}^i\}$, $i = 1, 2, \dots$; the sets T_j^i , $j = 1, \dots, j_i$, are, for each $i = 1, 2, \dots$, measurable and disjoint and $\bigcup_{j=1}^{j_i} T_j^i = T$; every element of P_τ^{i+1} is contained in some element of P_τ^i for $i = 1, 2, \dots$; and to every measurable subset E of T and every $\epsilon > 0$ there corresponds a positive integer $i(\epsilon)$ and a subset $J(E, \epsilon)$ of $\{1, 2, \dots, j_{i(\epsilon)}\}$ such that $|E - E_0| + |E_0 - E| < \epsilon$, where $E_0 = \bigcup_{j \in J(E, \epsilon)} T_j^{i(\epsilon)}$.

It is well known [9, Theorem C, p. 173] that there exists a dense sequence of partitions of T as a consequence of T being metric and compact, and the measure on T having the properties listed at the beginning of §2.

We shall require the following two known results.

LEMMA 4.2. *Let $\mathfrak{A} = L_T^1 \otimes C_R$ be the set of scalar functions ϕ on $T \times R$ such that $\phi(t, r) = \sum_{i=1}^k f_i(t) c_i(r)$, where k is some positive integer, $f_i \in L_T^1$, and $c_i \in C_R$, $i = 1, \dots, k$. Then \mathfrak{A} is a dense subset of \mathfrak{B} .*

Proof. [10, §6.4, p. 94].

LEMMA 4.3. *Let $l \in \mathfrak{B}^*$. Then there exists a measurable mapping ν from T to the class of regular signed Borel measures on R such that*

$$\langle l, \phi \rangle = \int_T \int_R \phi(t, r) \nu(dr; t) \quad \text{for all } \phi \in \mathfrak{B}$$

and

$$|\nu|(R; \cdot) \in L_T^\infty.$$

Proof. [11, Exposé no. 4, p. 3].

We shall also apply the following lemma.

LEMMA 4.4. Let $\epsilon > 0$, T_ϵ have the properties described in (i) of Assumption 2.3, F be a measurable subset of T_ϵ , $\{R_1, R_2, \dots, R_m\}$ be a partition of R into disjoint nonempty Borel subsets, and $\sigma \in \mathcal{S}$, and assume that the support of $\sigma(t)$ is contained in $\bar{R}^*(t)$ (the closure of $R^*(t)$) for all t in F . Let $\alpha^k = \int_F \sigma(R_k; t)$, $k = 1, \dots, m$. Then there exists a partition of F into disjoint measurable sets F_1, F_2, \dots, F_m and a measurable original control ρ such that, for $k = 1, 2, \dots, m$, $|F_k| = \alpha^k$, $\rho(t) \in R^*(t)$ on T , and $\rho(t)$ is within a distance 2ϵ of R_k a.e. in F_k .

Proof. Let k represent integers from 1 to m , and let $G_k = \{t \in F \mid \sigma(R_k; t) \neq 0\}$. For every nonempty subset A of $\{1, 2, \dots, m\}$, let $G_A = \bigcap_{k \notin A} (F - G_k) \bigcap_{k \in A} G_k$. Since $\sum_{k=1}^m \sigma(R_k; t) = \sigma(R; t) = 1$ in T , it follows that $F = \bigcup_A G_A$ (the union over all nonempty subsets A of $\{1, 2, \dots, m\}$). For every set A , we can partition G_A into disjoint measurable subsets G_A^k , $k \in A$, of measure $\int_{G_A} \sigma(R_k; t)$. If $k \notin A$, we define G_A^k to be the empty set. We now let $F_k = \bigcup_A G_A^k$, and verify that $|F_k| = \alpha^k$, $k = 1, \dots, m$.

Let k now be fixed. Since, by construction, $\sigma(R_k; t) \neq 0$ for $t \in F_k$, for every \bar{t} in F_k there exists a point r_i^k in $\bar{R}^*(\bar{t}) \cap R_k$ and, by (i) of Assumption 2.3, there exists a measurable original control ρ_i^k , continuous at \bar{t} when restricted to F , and such that $|\rho_i^k(\bar{t}), r_i^k| \leq \epsilon$ and $\rho_i^k(t) \in R^*(t)$ on T . Because ρ_i^k is continuous at \bar{t} when restricted to F , there exists (relative to F_k) a neighborhood $N_k(\bar{t})$ of \bar{t} such that $|\rho_i^k(t), r_i^k| < 2\epsilon$ for t in $N_k(\bar{t})$; hence $\rho_i^k(t)$ is within 2ϵ of R_k for $t \in N_k(\bar{t})$. Since F_k is covered by open (relative to F_k) neighborhoods $N_k(\bar{t})$, it must be covered a.e. by a denumerable subfamily, say $N_k(\bar{t}_1), N_k(\bar{t}_2), \dots$. We now let $\rho(t) = \rho_{i_j}^k(t)$ for $t \in N_k(\bar{t}_j) - \bigcup_{i=1}^{j-1} N_k(\bar{t}_i)$, $k = 1, \dots, m$, $j = 1, 2, \dots$, and $\rho(t) = \rho_{i_1}^1(t)$ everywhere else on T . Since $\rho_{i_j}^k \in \mathcal{R}^*$ for all k and j , it follows that $\rho(t) \in R^*(t)$ on T . We also observe that ρ is measurable and $\rho(t)$ is within a distance 2ϵ from R_k a.e. in F_k , $k = 1, 2, \dots, m$.

4.1. Proof of Theorem 2.4. Let $l \in \mathcal{R}^{**}$, and let a corresponding mapping ν be defined as in Lemma 4.3. Then $\nu(t)$ has its support in $\bar{R}^*(t)$ for every $t \in T$ and we can represent $\nu(t)$ as the difference of two bounded non-negative measures; specifically,

$$\nu(t) = h^1(t)\sigma_1(t) - h^2(t)\sigma_2(t) \quad \text{on } T,$$

where $\sigma_i \in \mathcal{S}^*$, $h^i \in L_T^\infty$, and $h^i(t) \geq 0$ on T , $i = 1, 2$.

We first consider those elements l in \mathcal{R}^{**} for which $h^1(t) + h^2(t) = 1$ on T . Then,

$$\sigma^+ = h^1\sigma_1 + h^2\sigma_2 \in \mathcal{S}^*.$$

Let the sets $T_j^i, j = 1, \dots, j_i, i = 1, 2, \dots$, define a dense sequence of partitions of T as in Definition 4.1. Since R is metric and compact, for every positive integer i we can partition R into disjoint Borel subsets $R_k^i, k = 1, \dots, k_i$, of diameters not exceeding $1/i$. In every one of these sets R_k^i we may arbitrarily select a point r_k^i . Let $T_{1/i}$ be defined as in Assumption 2.3, and let $T_j^{*i} = T_j^i \cap T_{1/i}, j = 1, 2, \dots, j_i, i = 1, 2, \dots$. For every positive integer i and for every $j = 1, 2, \dots, j_i$, we may define sets $T_{j,k}^i, k = 1, \dots, k_i$, and controls $\rho_j^i \in \mathfrak{R}^*$ that have the properties described in the statement of Lemma 4.3, with $\sigma^+, 1/i, T_j^{*i}, T_{j,k}^i, R_k^i$, and ρ_j^i replacing $\sigma, \epsilon, F, F_k, R_k$, and ρ , respectively. Let now a measurable original control ρ_i be defined for $i = 1, 2, \dots$ by the relations

$$\rho_i(t) = \begin{cases} \rho_j^i(t) & \text{on } T_j^{*i}, \quad j = 1, \dots, j_i, \\ \rho_1^1(t) & \text{on } T - T_{1/i}. \end{cases}$$

We observe that $\rho_i(t) \in R^*(t)$ on $T, |\rho_i(t), r_k^i| \leq 3/i$ a.e. in $T_{j,k}^i, j = 1, \dots, j_i$, and $|T_{j,k}^i| = \int_{T_j^{*i}} \sigma^+(R_k^i; t), k = 1, 2, \dots, k_i$. Let now, for $i = 1, 2, \dots, j = 1, \dots, j_i, k = 1, \dots, k_i, m = 1, 2$,

$$\theta_{j,k}^{i,m} = \begin{cases} \frac{1}{|T_{j,k}^i|} \int_{T_j^{*i}} h^m(t) \sigma_m(R_k^i; t) & \text{if } |T_{j,k}^i| \neq 0, \\ \frac{1}{2}(1 + (-1)^m) & \text{if } |T_{j,k}^i| = 0. \end{cases}$$

Then $\theta_{j,k}^{i,1} + \theta_{j,k}^{i,2} = 1$, and we can partition each set $T_{j,k}^i$ into two measurable subsets $T_{j,k}^{i,m}$ such that $|T_{j,k}^{i,m}| = \theta_{j,k}^{i,m} |T_{j,k}^i|, m = 1, 2$. We now define functions h_i^m and h_i for $m = 1, 2, i = 1, 2, \dots$, by

$$h_i^m(t) = \begin{cases} (-1)^{m+1} & \text{on } T_{j,k}^{i,m}, \quad j = 1, 2, \dots, j_i, \quad k = 1, 2, \dots, k_i, \\ \frac{1}{2} & \text{on } T - T_{1/i}, \\ 0 & \text{elsewhere,} \end{cases}$$

$$h_i(t) = h_i^1(t) + h_i^2(t).$$

Let now $\epsilon > 0, c \in C_R(|c| = \sup_{r \in R} |c(r)|)$, and let E be a measurable subset of T . We may choose an integer i_0 sufficiently large so that, for every $i \geq i_0$, there exists a subset J_i of $\{1, 2, \dots, j_i\}$ and a measurable set E_i in T such that $E_i = \bigcup_{j \in J_i} T_j^{*i}, |E - E_i| + |E_i - E| < \epsilon/(8|c|)$ and $|c(r) - c(r')| < \epsilon/(8|T|)$ if $|r, r'| \leq 3/i$. Finally, let $O(\alpha)$ represent here a quantity whose absolute value does not exceed α . Then, for $i \geq i_0$,

$$\int_E \int_R c(r) \nu(dr; t) = \sum_{m=1}^2 (-1)^{m+1} \int_E h^m(t) \int_R c(r) \sigma_m(dr; t)$$

$$\begin{aligned}
&= \sum_{m=1}^2 (-1)^{m+1} \int_E h^m(t) \sum_{k=1}^{k_i} c(r_k^i) \sigma_m(R_k^i; t) + O(\epsilon/4) \\
&= \sum_{m=1}^2 (-1)^{m+1} \sum_{j \in J_i} \sum_{k=1}^{k_i} c(r_k^i) \int_{T_{j,*}^i} h^m(t) \sigma_m(R_k^i; t) + O(\epsilon/2) \\
&= \sum_{m=1}^2 (-1)^{m+1} \sum_{j \in J_i} \sum_{k=1}^{k_i} c(r_k^i) |T_{j,k}^{i,m}| + O(\epsilon/2) \\
&= \sum_{m=1}^2 \sum_{j \in J_i} \sum_{k=1}^{k_i} \int_{T_{j,k}^{i,m}} c(r_k^i) h_i^m(t) + O(\epsilon/2) \\
&= \sum_{j \in J_i} \sum_{k=1}^{k_i} \sum_{m=1}^2 \int_{T_{j,k}^{i,m}} h_i^m(t) c(\rho_i(t)) + O(3\epsilon/4) \\
&= \sum_{j \in J_i} \int_{T_{j,*}^i} h_i(t) c(\rho_i(t)) + O(3\epsilon/4) \\
&= \int_E h_i(t) c(\rho_i(t)) + O(\epsilon).
\end{aligned}$$

Thus,

$$\lim_{i \rightarrow \infty} \int_T f_E(t) h_i(t) c(\rho_i(t)) = \int_T \int_R c(r) f_E(t) \nu(dr; t)$$

for every $c \in C_R$ and every measurable characteristic function f_E on T , provided $h^1(t) + h^2(t) = 1$ on T . It follows that

$$\lim_{i \rightarrow \infty} \int_T f(t) h_i(t) c(\rho_i(t)) = \int_T \int_R c(r) f(t) \nu(dr; t)$$

for every $c \in C_R$ and $f \in L_T^1$. Since every ν of Lemma 4.3 is such that $\nu(t) = h(t)(h^1(t)\sigma_1(t) - h^2(t)\sigma_2(t))$ on T , where $h \in L_T^\infty$, $h^m(t) \geq 0$, $m = 1, 2$, and $h^1(t) + h^2(t) = 1$ on T , it follows that

$$\lim_{i \rightarrow \infty} \int_T f(t) \bar{h}_i(t) c(\rho_i(t)) = \int_T \int_R c(r) f(t) \nu(dr; t),$$

where $\bar{h}_i(t) = h(t)h_i(t)$ on T . Now, by Lemma 4.2, every $\phi \in \mathfrak{B}$ can be approximated in \mathfrak{B} by finite sums of terms of the form $c(r)f(t)$ ($c \in C_R$, $f \in L_T^1$), and we conclude that

$$\lim_{i \rightarrow \infty} \int_T \bar{h}_i(t) \phi(t, \rho_i(t)) = \int_T \int_R \phi(t, r) \nu(dr; t)$$

for every $\phi \in \mathfrak{B}$ and every $\nu \in \mathfrak{B}^{**}$. Thus \mathfrak{B}^{**} is the closure of $L_T^\infty \times \mathfrak{B}^*$.

We observe that, by construction, $\bar{h}_i(t) = h_i(t) = 1$ a.e. in T if $\nu \in \mathfrak{S}^*$. It follows that \mathfrak{S}^* is the closure of \mathfrak{B}^* .

This completes the proof of the theorem.

4.2. Proof of Theorem 2.5. We first observe that \mathcal{S} is sequentially compact. Indeed, for every $\phi \in \mathcal{B}$,

$$|\langle \sigma, \phi \rangle| = \left| \int_T \int_R \phi(t, r) \sigma(dr; t) \right| \leq \|\phi\|;$$

hence the \mathcal{B}^* -norm of σ is ≤ 1 , and every sequence in \mathcal{S} has a subsequence converging to some point in \mathcal{B}^* .

Let $\nu \in \mathcal{B}^*$ be a limit of a sequence $\sigma_1, \sigma_2, \dots$ in \mathcal{S} . Then $\langle \nu, \phi \rangle \geq 0$ for every nonnegative ϕ in \mathcal{B} , and $\langle \nu, \bar{\phi} \rangle = |E|$ if $\bar{\phi}(t, r) = 1$ on $E \times R$ and $\bar{\phi}(t, r) = 0$ on $(T - E) \times R$ for every measurable $E \subset T$. Since $\nu(t)$ is a regular Borel measure for every t , it follows that $\nu(t) \in S$ a.e. in T and, therefore, ν is equivalent to an element of \mathcal{S} .

We shall next show that if a sequence $\sigma_1, \sigma_2, \dots$ converges to σ , and if $\sigma_j(\bar{R}^*(t); t) = 1$ a.e. in T , $j = 1, 2, \dots$, then $\sigma(\bar{R}^*(t); t) = 1$ a.e. in T . Indeed, let $\eta > 0$ and let T_η , of measure at least $|T| - \eta$, be a closed subset of T such that R^* is continuous when restricted to T_η . Let $\epsilon > 0$, $\bar{t} \in T_\eta$, $S_\eta(\bar{t}, \delta) = \{t \in T_\eta \mid |t, \bar{t}| < \delta\}$, $U(R', h)$ be the open h -neighborhood of a set $R' \subset R$, $U_{1/2} = U(\bar{R}^*(\bar{t}), \epsilon/2)$, $\bar{U}_{1/2}$ be the closure of $U_{1/2}$, $U_1 = U(\bar{R}^*(\bar{t}), \epsilon)$, and let $\delta = \delta(\epsilon)$ be such that $\bar{R}^*(t) \subset U_{1/2}$ and $\bar{R}^*(\bar{t}) \subset U(\bar{R}^*(t), \epsilon/2)$ for all t in $S_\eta(\bar{t}, \delta)$. Let \bar{c} in C_R be such that $\bar{c}(r) = 0$ on $\bar{U}_{1/2}$, $0 \leq \bar{c}(r) \leq 1$ on R , and $\bar{c}(r) = 1$ on $R - U_1$. Then,

$$\begin{aligned} 0 &= \lim_{j \rightarrow \infty} \int_{S_\eta(\bar{t}, \delta)} \int_R \bar{c}(r) \sigma_j(dr; t) = \int_{S_\eta(\bar{t}, \delta)} \int_R \bar{c}(r) \sigma(dr; t) \\ &\geq \int_{S_\eta(\bar{t}, \delta)} \sigma(R - U_1; t); \end{aligned}$$

hence, $\sigma(R - U_1; t) = 0$ a.e. in $S_\eta(\bar{t}, \delta(\epsilon))$ or $\sigma(U_1; t) = 1$ a.e. in $S_\eta(\bar{t}, \delta(\epsilon))$. We observe that, for all t in $S_\eta(\bar{t}, \delta(\epsilon))$, $U_1 \subset U(\bar{R}^*(t), 3\epsilon/2)$. It follows that T_η can be covered by open (relative to T_η) neighborhoods in each of which $\sigma(t)$ is a.e. supported by $U(\bar{R}^*(t), 3\epsilon/2)$. Since ϵ is arbitrary, $\bar{R}^*(t)$ is compact for all t , and the measure $\sigma(t)$ is regular, we conclude that $\sigma(\bar{R}^*(t); t) = 1$ a.e. in T_η . Since η is arbitrary, it follows that $\sigma(\bar{R}^*(t); t) = 1$ a.e. in T . Thus \mathcal{S}^* is sequentially compact.

It follows from the above conclusion and from the continuity of (y^0, y) on the sequentially compact space $\mathcal{S}^* \times B_0$ that the set \hat{Y} is closed. By Theorem 2.4, \hat{Y} is contained in the closure of \hat{X} . Since \hat{X} is obviously a subset of \hat{Y} , we conclude that \hat{Y} is the closure of \hat{X} . This completes the proof of the theorem.

4.3. Proof of Theorem 2.7. Since ψ is measurable on T , and in view of Lemma 4.2 and of Assumption 2.3, for every positive integer i there exists a compact subset F_i of T , of measure at least $|T| - 1/i$, such that ψ and R^* are continuous, hence uniformly continuous, when restricted to F_i , and ϕ

can be approximated on $F_i \times R$ to within $1/(3i)$ by a uniformly continuous function.

Let $i \geq 1$, and let $\delta = \delta(i)$ be such that $|\psi(t) - \psi(t')| \leq 1/i$, $|\phi(t, r) - \phi(t', r)| \leq 1/i$ on R , and $R^{\#}(t) \subset U(R^{\#}(t'), 1/i)$ for $|t, t'| \leq \delta$, $t \in F_i$, $t' \in F_i$. We can cover F_i by a finite number of measurable sets $F_i^1, F_i^2, \dots, F_i^k$ of diameters less than δ . Let, for $j = 1, \dots, k$,

$$D_i^j = F_i^j - \bigcup_{m=1}^{j-1} D_i^m, \quad D_i^0 = T - F_i,$$

and let, for $j = 0, 1, \dots, k$, $t_j \in D_i^j$, χ_j^i be the characteristic function of D_i^j , $\sigma_i(t) = \sum_{j=0}^k \chi_j^i(t) \bar{\sigma}(t_j)$, and $\psi_i(t) = \int_R \phi(t, r) \sigma_i(dr; t)$ on T . Then σ_i and ψ_i are measurable, $\sigma_i(U(\bar{R}^{\#}(t), 1/i); t) = 1$ on F_i , and

$$\begin{aligned} |\psi(t) - \psi_i(t)| &\leq |\psi(t) - \psi(t_j)| + |\psi(t_j) - \psi_i(t_j)| \\ &\leq 2/i \quad \text{for } t \in D_i^j, \quad j = 1, \dots, k, \end{aligned}$$

hence on F_i . Thus the ψ_i converge in measure to ψ as $i \rightarrow \infty$, and there exists a sequence J_1 of positive integers such that $\psi_i \rightarrow \psi$ a.e. for $i \in J_1$, $i \rightarrow \infty$. By Theorem 2.5, there exists a subsequence J_2 of J_1 such that the sequence $\{\sigma_i\}_{i \in J_2}$ converges to some $\bar{\sigma} \in \mathcal{S}$. We let $\bar{\psi}(t) = \int_R \phi(t, r) \bar{\sigma}(dr; t)$ on T .

For every measurable set E in T ,

$$\lim_{i \in J_2} \int_E \psi_i(t) = \lim_{i \in J_2} \int_E \int_R \phi(t, r) \sigma_i(dr; t) = \int_E \int_R \phi(t, r) \bar{\sigma}(dr; t) = \int_E \bar{\psi}(t).$$

Since $\psi_i \rightarrow \psi$ a.e. for $i \in J_2$, and $|\psi_i(t)| \leq \sup_{r \in R} |\phi(t, r)|$, it follows that $\lim_{i \in J_2} \int_E \psi_i(t) = \int_E \psi(t)$; hence $\int_E \psi(t) = \int_E \bar{\psi}(t)$ for every measurable E in T . We conclude that $\psi(t) = \bar{\psi}(t) = \int_R \phi(t, r) \bar{\sigma}(dr; t)$ a.e. in T .

We shall next verify that $\bar{\sigma}(\bar{R}^{\#}(t); t) = 1$ a.e. in T . Indeed, let $\epsilon > 0$, and let $R_\epsilon^{\#}(t) = U(\bar{R}^{\#}(t), \epsilon)$. We observe that $\sigma_i(\bar{R}_\epsilon^{\#}(t); t) = \bar{\sigma}(\bar{R}_\epsilon^{\#}(t); t_j) = 1$ on D_i^j for all $i \geq 1/\epsilon$ and all corresponding j . Since $|F_i| \geq |T| - 1/i$, and since the mapping $R_\epsilon^{\#}$ satisfies (ii) of Assumption 2.3 if $R^{\#}$ does, it follows from the argument of §4.2 that $\sigma(\bar{R}_\epsilon^{\#}(t); t) = 1$ a.e. in T for every $\epsilon > 0$. Since $\bar{R}^{\#}(t)$ is compact (as a closed subset of a compact set), it follows that $\bar{\sigma}(\bar{R}^{\#}(t); t) = 1$ a.e. in T .

We may now choose a mapping σ that coincides with $\bar{\sigma}$ wherever $\bar{\sigma}(\bar{R}^{\#}(t); t) = 1$ and $\psi(t) = \bar{\psi}(t)$ and that equals $\bar{\sigma}$ elsewhere. This mapping satisfies the statement of the theorem.

5. Original and relaxed curves. We shall now apply Theorems 2.4, 2.5, and 2.7 to prove Theorem 3.1.

Let $\sigma \in \mathcal{S}^{\#}$, $b \in B_0$, and let $\xi = \xi(\cdot; \sigma, b)$ be an absolutely continuous curve such that

$$(5.1) \quad d\xi(t)/dt = \xi(t) = \int_R g(\xi(t), t, r) \sigma(dr; t) \quad \text{a.e. in } T,$$

$$(5.2) \quad \xi(t_0) = b.$$

Condition (3.1) and the assumption that $g^i(v, \cdot, \cdot) \in \mathcal{B}$, $v \in E_n$, $i = 1, \dots, n$, imply that

$$(5.3) \quad \begin{aligned} & |g(v, t, r)| \\ & \leq |g(b_0, t, r)| + \psi(t)|v - b_0| \leq g_{\sup}(b_0, t) + \psi(t)|v - b_0| \\ & \leq \psi_0(t)(1 + |v|) \quad \text{on } E_n \times T \times R, \end{aligned}$$

where b_0 is fixed in B_0 and ψ , $g_{\sup}(b_0, \cdot)$ and ψ_0 are integrable on T . It follows then, by well-known existence theorems, that such a solution ξ exists and ξ is bounded by an integrable ψ_1 for all $\sigma \in \mathcal{S}^{\#}$ and $b \in B_0$; furthermore, if $b_i \rightarrow b$ as $i \rightarrow \infty$, $\xi(t; \sigma, b_i)$ converges to $\xi(t; b, \sigma)$, uniformly on $T \times \mathcal{S}^{\#}$.

We shall now prove that, for a fixed $b \in B_0$, $\xi(\cdot; \sigma_j, b)$ converges uniformly to $\xi(\cdot; \sigma, b)$ if $\sigma_j \rightarrow \sigma$ in $\mathcal{S}^{\#}$. Assume that $\sigma_j \rightarrow \sigma$ in $\mathcal{S}^{\#}$, let b be fixed, $\xi = \xi(\cdot; \sigma, b)$, $\xi_j = \xi(\cdot; \sigma_j, b)$, and $\eta_j = \xi - \xi_j$, $j = 1, 2, \dots$. Then for $\tau \in T$ and $j = 1, 2, \dots$,

$$\begin{aligned} \eta_j(\tau) &= \int_{t_0}^{\tau} dt \int_R g(\xi(t), t, r) \sigma(dr; t) - \int_{t_0}^{\tau} dt \int_R g(\xi_j(t), t, r) \sigma_j(dr; t) \\ &= \int_{t_0}^{\tau} dt \int_R g(\xi(t), t, r) (\sigma(dr; t) - \sigma_j(dr; t)) \\ &\quad + \int_{t_0}^{\tau} dt \int_R (g(\xi(t), t, r) - g(\xi_j(t), t, r)) \sigma_j(dr; t). \end{aligned}$$

Let now

$$h_j(\tau) = \left| \int_{t_0}^{\tau} dt \int_R g(\xi(t), t, r) (\sigma(dr; t) - \sigma_j(dr; t)) \right|.$$

Then

$$(5.4) \quad |\eta_j(\tau)| \leq h_j(\tau) + \int_{t_0}^{\tau} \psi(t) |\eta_j(t)| dt.$$

We now observe that the function γ , defined by $\gamma(t, r) = g(\xi(t), t, r)$ on $T \times R$, belongs to $\mathcal{B} \times \dots \times \mathcal{B}$. It follows that $h_j(\tau) \rightarrow 0$ on T as

$j \rightarrow \infty$, while, by (5.3), $0 \leq h_j(\tau) \leq 2 \int_{t_0}^{t_1} \psi_0(t)(1 + |\xi(t)|) dt < \infty$. Let now

$$v_j(\tau) = \int_{t_0}^{\tau} \psi(t) |\eta_j(t)| dt \quad \text{and} \quad w(\tau) = \int_{t_0}^{\tau} \psi(t) dt.$$

Then, by (5.4),

$$e^{-w(\tau)}(\dot{v}_j(\tau) - \dot{w}(\tau)v_j(\tau)) \leq e^{-w(\tau)}\psi(\tau)h_j(\tau) \quad \text{on } T$$

and

$$e^{-w(\tau)}v_j(\tau) \leq \int_{t_0}^{\tau} e^{-w(t)}\psi(t)h_j(t) dt \leq \int_{t_0}^{t_1} \psi(t)h_j(t) dt.$$

It follows now from the Lebesgue dominated convergence theorem that $\int_{t_0}^{t_1} \psi(t)h_j(t) dt \rightarrow 0$ as $j \rightarrow \infty$; hence $v_j(\tau) \rightarrow 0$ and, by (5.4), $\xi_j(\tau) \rightarrow \xi(\tau)$ for every τ in T . Since the ξ_j are equicontinuous, it follows that $\xi_j \rightarrow \xi$ uniformly on T .

We now consider an arbitrary relaxed curve η and recall that, by definition, $\dot{\eta}(t) \in F^*(\eta(t), t)$ a.e. in T . Now, every point in $F^*(\eta(t), t)$ can be represented as $\int_R g(\eta(t), t, r)\bar{\sigma}(dr; t)$, where $\bar{\sigma}(\bar{R}^*(t); t) = 1$ and, by Theorem 2.7, we may assume that $\bar{\sigma} \in S^*$. Thus, $\eta = \xi(\cdot; \bar{\sigma}, \eta(t_0))$ and every relaxed curve η is such that $\dot{\eta}(t) = \int_R g(\eta(t), t, r)\bar{\sigma}(dr; t)$ a.e. in T for some $\bar{\sigma} \in S^*$. Furthermore, given a sequence ξ_1, ξ_2, \dots of relaxed curves, corresponding to some relaxed controls $\sigma_1, \sigma_2, \dots$ in S^* and some initial points $b_1 = \xi_1(t_0), b_2 = \xi_2(t_0), \dots$ in B_0 , there exist, by Theorem 2.5, a relaxed control $\sigma \in S^*$, a point $b \in B_0$, and a sequence $J = \{j_1, j_2, \dots\}$ of positive integers such that $\sigma_i \rightarrow \sigma$ and $b_i \rightarrow b$ as $i \rightarrow \infty, i \in J$. Let now $\xi(\cdot; \sigma, b)$ be defined on T by system (5.1), (5.2). Then, for every $t \in T$,

$$\begin{aligned} |\xi(t; \sigma, b) - \xi_i(t)| &= |\xi(t; \sigma, b) - \xi(t; \sigma_i, b_i)| \\ &\leq |\xi(t; \sigma, b) - \xi(t; \sigma_i, b)| + |\xi(t; \sigma_i, b) - \xi(t; \sigma_i, b_i)|. \end{aligned}$$

We have proven that $\xi(t; \sigma_i, b) \rightarrow \xi(t; \sigma, b)$, and we have observed previously that $|\xi(t; \sigma_i, b) - \xi(t; \sigma_i, b_i)| \rightarrow 0$ as $i \rightarrow \infty$ (because of the continuity in b , uniformly on S). It follows that $\xi_i(t) \rightarrow \xi(t; \sigma, b)$ as $i \rightarrow \infty, i \in J$, for every $t \in T$, hence (as previously observed) uniformly on T . Since $\xi_i(t) \in A, t \in T, i = 1, 2, \dots, \xi_i(t_0) \in B_0, i = 1, 2, \dots$, and A and B_0 are closed, it follows that $\xi(t) \in A, t \in T$ and $\xi(t_0) \in B_0$. Thus ξ is a relaxed curve, and this proves that the set \tilde{Y} is compact in the topology of the uniform convergence.

We next consider a relaxed curve ξ and the corresponding relaxed control $\sigma \in \mathcal{S}^*$ and initial point $b = \xi(t_0) \in B_0$. By Theorem 2.4, the relaxed control σ can be approximated by original controls ρ_1, ρ_2, \dots in \mathcal{R}^* . We let $\xi_j = \xi(\cdot; \rho_j, b)$, $j = 1, 2, \dots$, and observe that, since $\xi(\cdot, b)$ is continuous on \mathcal{S}^* , $\xi_j \rightarrow \xi(\cdot; \sigma, b) = \xi$ uniformly. This completes the proof of the theorem.

Acknowledgment. The author wishes to express his appreciation to Professor R. Bonic for helpful advice and stimulating discussions.

REFERENCES

- [1] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111-128.
- [2] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129-145.
- [3] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Soc. Sci. Lettres Varsovie, CL III, 30 (1937), pp. 212-234.
- [4] ———, *Necessary conditions in the calculus of variations*, Acta Math., 69 (1938), pp. 239-258.
- [5] J. WARGA, *On a class of minimax problems in the calculus of variations*, Michigan Math J., 12 (1965), pp. 289-311.
- [6] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Part I, Interscience, New York, 1964.
- [7] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, this Journal, 1 (1962), pp. 76-84.
- [8] E. ROXIN, *The existence of optimal controls*, Michigan Math J., 9 (1962), pp. 109-119.
- [9] P. R. HALMOS, *Measure Theory*, Van Nostrand, New York, 1950.
- [10] H. H. SCHAEFER, *Topological Vector Spaces*, Macmillan, New York, 1966.
- [11] L. SCHWARTZ, *Produits tensoriels topologiques d'espaces vectoriels topologiques. Espaces vectoriels topologiques nucléaires. Applications*, Séminaire 1953/1954, Faculté des Sciences de Paris, 1954.
- [12] E. J. MCSHANE AND R. B. WARFIELD, JR., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41-47.
- [13] C. CASTAING, *Quelques problèmes de mesurabilité liés à la théorie de la commande*, C. R. Acad. Sci. Paris, 262 (1966), pp. 409-411.

RESTRICTED MINIMA OF FUNCTIONS OF CONTROLS*

J. WARGA†

1. Introduction. In a previous paper [1] we have discussed the existence of controls ρ that minimize a function x^0 subject to the restrictions that, for every value of its argument t in a metric space, $\rho(t)$ is contained in some preassigned set $R^\#(t)$ and that $x(\rho) \in B_1$, where x is a given mapping and B_1 is a closed subset of a topological space H . We have shown that, in a large class of problems, such minimizing controls exist in a larger space of "relaxed controls" and that these relaxed controls can be approximated by original controls.

In this paper we shall assume that H is the Euclidean n -space E_n . We wish to investigate certain necessary conditions for minimum that might be considered a generalization of the Weierstrass E -condition and of the transversality conditions of the calculus of variations. In this sense our results represent an extension of certain methods and theorems of the mathematical control theory, and specifically of [2] and [3], to a more general setting. The necessary conditions that we obtain are no longer restricted, however, to minima over the space \mathcal{S} of relaxed controls but apply as well to minima over the space \mathcal{R} of original controls (if such minima exist). Thus our present results also generalize Pontryagin's Maximum Principle. Furthermore, the space \mathcal{R} is no longer restricted, as in [1], to measurable mappings from one metric compact set to another.

Previous attempts to apply the methods of the mathematical control theory to problems involving functions defined otherwise than by a system of differential or difference equations were mostly limited to special, and linear, problems. Recent results of Neustadt [4], [5] are, however, quite general. They are based on a separation theorem for convex sets that represent certain linearizations of constraints. Our approach is, however, different from Neustadt's; in particular, our basic results are stated directly in the form of inequalities involving the value of the minimizing control at an arbitrary point rather than in the form of functional inequalities.

Let T and R be arbitrary sets, B a convex set, \mathcal{R} a class of controls, that is, mappings from T to R , $x = (x^1, \dots, x^n)$ a given function from $\mathcal{R} \times B$ to E_n , and B_1 a given set in E_n . We wish to characterize a control $\bar{\rho} \in \mathcal{R}$ and a point $\bar{b} \in B$ that yield a minimum of $x^1(\rho, b)$ subject to the condition $x(\rho, b) \in B_1$. The necessary conditions for minimum that we derive are

* Received by the editors May 22, 1967, and in final revised form August 23, 1967.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115. This research was supported by the National Aeronautics and Space Administration under Grant NGR 22-011-020.

expressed in terms of certain variational derivatives $Dx(\bar{p}, \bar{b}; t^*, \rho^*)$, respectively $Dx(\bar{p}, \bar{b}; t^*, r)$, defined in §2. These derivatives represent, roughly speaking, the rate of change of x when its argument \bar{p} is replaced by the function ρ^* (respectively the constant function r) over a "small" set in the "neighborhood" of t^* .

As an illustration, we consider, in §3 and §5, the "standard" problem of the mathematical control theory of ordinary differential equations and prove a slight generalization of the usual necessary conditions.

2. Necessary conditions for minimum. Let T and R be arbitrary sets, B a convex set, \mathfrak{R} an arbitrary class of mappings from T to R , and B_1 a set in E_n . The vector function $x = (x^1, \dots, x^n)$ is a given mapping from $\mathfrak{R} \times B$ to E_n . If $\rho: T \rightarrow R$, we denote by $\rho(t)$ the image under the mapping ρ of a point t in T . If the mapping ρ depends on some parameters a, b, c , we designate by $\rho(a, b, c)$, or by $\rho(\cdot; a, b, c)$, the mapping, and by $\rho(t; a, b, c)$ the image of t under the mapping. Similarly, x denotes a mapping and $x(\rho)$ the image of ρ under the mapping x . We also write, when it appears more appropriate, $t \rightarrow \rho(t)$ to represent a mapping.

If ρ_1 and ρ_2 are two mappings from T to R , and A is a subset of T , we designate by $[\rho_1, A; \rho_2]$ the mapping ρ defined by the relations

$$\rho(t) = \begin{cases} \rho_1(t) & \text{on } A, \\ \rho_2(t) & \text{on } T - A. \end{cases}$$

Similarly, if $\rho_1, \rho_2, \dots, \rho_k, \bar{p}$ are mappings from T to R , and A_1, A_2, \dots, A_k are disjoint subsets of T , we designate by $[\rho_i, A_i (i = 1, \dots, k); \bar{p}]$ the mapping ρ defined by the relations

$$\rho(t) = \begin{cases} \rho_i(t) & \text{on } A_i, \quad i = 1, \dots, k, \\ \bar{p}(t) & \text{elsewhere on } T. \end{cases}$$

Let T^* be a subset of T , and let \mathfrak{M} be a collection of subsets $M(t, \alpha)$ of T , $t \in T^*$, $\alpha \geq 0$. Let $\bar{p} \in \mathfrak{R}$, $\rho^* \in \mathfrak{R}$, $\bar{b} \in B$, $t^* \in T$, $\alpha > 0$, and let $\rho' = [\rho^*, M(t^*, \alpha); \bar{p}]$.

If $\rho' \in \mathfrak{R}$ for sufficiently small α , and if

$$\lim_{\alpha \rightarrow +0} \frac{1}{\alpha} (x(\rho', \bar{b}) - x(\bar{p}, \bar{b}))$$

exists, we shall say that " x has an \mathfrak{M} -derivative at (\bar{p}, \bar{b}) with respect to (t^*, ρ^*) " and we shall designate this limit by $D_{\mathfrak{M}}x(\bar{p}, \bar{b}; t^*, \rho^*)$. If $\mathfrak{R}^*(t^*)$ is a subset of \mathfrak{R} for each $t^* \in T^*$ and $D_{\mathfrak{M}}x(\bar{p}, \bar{b}; t^*, \rho^*)$ is the same for all $\rho^* \in \mathfrak{R}^*(t^*)$ such that $\rho^*(t^*) = r$, we shall write $D_{\mathfrak{M}, \mathfrak{R}^*}x(\bar{p}, \bar{b}; t^*, r)$ or $Dx(\bar{p}, \bar{b}; t^*, r)$ (if \mathfrak{M} and the mapping $t^* \rightarrow \mathfrak{R}^*(t^*)$ are fixed).

Let now $\bar{p} \in \mathfrak{R}$, $\bar{b} \in B$, and $b \in B$. We shall write $Dx(\bar{p}, \bar{b}; b)$ for $\lim_{\theta \rightarrow +0} (1/\theta)(x(\bar{p}, (1 - \theta)\bar{b} + \theta b) - x(\bar{p}, \bar{b}))$.

DEFINITION 2.1. Let $\bar{p} \in \mathcal{R}$, $\bar{b} \in B$, $T^* \subset T$, let \mathfrak{K}_k be, for $k = 1, 2, \dots, n^2$, a collection of subsets $N_k(t, \alpha)$ of T , $t \in T^*$, $\alpha \geq 0$, and let $\mathfrak{K} = \{\mathfrak{K}_k \mid k = 1, \dots, n^2\}$. Let \mathcal{R}^* be a mapping from T^* to the class of subsets of \mathcal{R} . We shall say that $(T^*, \mathcal{R}^*, \mathfrak{K})$ defines *local variations for x in $\mathcal{R} \times B$ at (\bar{p}, \bar{b})* if the following conditions hold:

- (i) For $t^*, t_1^*, t_2^* \in T^*$, $k, k_1, k_2 = 1, 2, \dots, n^2$, and $\alpha, \beta \geq 0$, $N_k(t^*, \alpha) \subset N_k(t^*, \beta)$ if $\alpha \leq \beta$; $N_k(t^*, 0)$ is the empty set; $N_{k_1}(t^*, \alpha)$ and $N_{k_2}(t^*, \beta)$ are disjoint if $k_1 \neq k_2$; and $N_{k_1}(t_1^*, \alpha)$ and $N_{k_2}(t_2^*, \beta)$ are disjoint if $t_1^* \neq t_2^*$ and both α and β are sufficiently small.
- (ii) Let an array with elements β^{ij} , $i, j = 1, \dots, n$, be represented by β^\square . For every choice of t^\square with elements $t^{ij} \in T^*$ and of ρ^\square with elements $\rho^{ij} \in \mathcal{R}^*(t^{ij})$, let the set $\Omega = \Omega(t^\square)$ in E_n contain all arrays ω^\square with $\omega^{ij} \geq 0$ which are such that the sets $N_{nj-n+i}(t^{ij}, \omega^{ij})$, $i, j = 1, \dots, n$, are disjoint, and let $\rho' = \rho'(t^\square, \rho^\square, \omega^\square) = [\rho^{ij}, N_{nj-n+i}(t^{ij}, \omega^{ij}) (i, j = 1, \dots, n); \bar{p}]$ for $\omega^\square \in \Omega(t^\square)$. Finally, let b^\square have elements $b^{ij} \in B$, 0^\square have elements $0^{ij} = 0$, $\mathfrak{I} = \{\theta^\square \mid \theta^{ij} \geq 0, \sum_{i,j=1}^n \theta^{ij} \leq 1\}$, $\theta^0 = 1 - \sum_{i,j=1}^n \theta^{ij}$, and $\theta^0 \circ b^\square = \theta^0 \bar{b} + \sum_{i,j=1}^n \theta^{ij} b^{ij}$.

Then:

- (a) $\rho' \in \mathcal{R}$;
- (b) for fixed t^\square, ρ^\square and b^\square , the function $(\omega^\square, \theta^\square) \rightarrow \xi(\omega^\square, \theta^\square) = \xi(\omega^\square, \theta^\square, t^\square, \rho^\square, b^\square) = x(\rho'(t^\square, \rho^\square, \omega^\square), \theta^\square \circ b^\square)$ from $\Omega \times \mathfrak{I}$ to E_n is continuous in some neighborhood of $(0^\square, 0^\square)$, and has a differential at $(0^\square, 0^\square)$ (relative to $\Omega \times \mathfrak{I}$);
- (c) for every $t^* \in T^*$, $\rho^* \in \mathcal{R}^*(t^*)$, and $k = 1, 2, \dots, n^2$, $D_{\mathfrak{K}_k} x(\bar{p}, \bar{b}; t^*, \rho^*) = Dx(\bar{p}, \bar{b}; t^*, r)$ exists, is independent of k , and has the same value for all $\rho^* \in \mathcal{R}^*(t^*)$ such that $\rho^*(t^*) = r$.

We can now state our general necessary conditions for minimum which we shall prove in §4.

THEOREM 2.2. Let (\bar{p}, \bar{b}) yield the minimum of $x^1(\rho, b)$ in $\mathcal{R} \times B$ subject to the condition $x(\rho, b) \in B_1$. Let $(T^*, \mathcal{R}^*, \mathfrak{K})$ define local variations for x in $\mathcal{R} \times B$ at (\bar{p}, \bar{b}) , and let, for all $t^* \in T^*$, $R^*(t^*) = \{\rho^*(t^*) \mid \rho^* \in \mathcal{R}^*(t^*)\}$. Let, furthermore, B_1^* be a convex set in E_m , \bar{b}_1^* a point in B_1^* , and $\phi: B_1^* \rightarrow B_1$ a continuous mapping such that $\phi(\bar{b}_1^*) = x(\bar{p}, \bar{b})$, $\phi(B_1^*) \subset B_1$ and ϕ has a differential at \bar{b}_1^* (relative to B_1^*), $d\phi(\bar{b}_1^*; b_1^* - \bar{b}_1^*) = \phi'(\bar{b}_1^*)(b_1^* - \bar{b}_1^*)$ (where $\phi'(\bar{b}_1^*)$ is a linear operator from E_m to E_n). Then, either

$$(2.1) \quad \phi'(\bar{b}_1^*)\bar{b}_1^* = \min_{b_1^* \in B_1^*} \phi'(\bar{b}_1^*)b_1^*,$$

or there exists a nonvanishing vector λ in E_n such that

$$(2.2) \quad \lambda \cdot Dx(\bar{p}, \bar{b}; t^*, r) \geq 0 \quad \text{for all } t^* \in T^* \quad \text{and } r \in R^*(t^*);$$

$$(2.3) \quad \lambda \cdot Dx(\bar{p}, \bar{b}; b) \geq 0 \quad \text{for all } b \in B;$$

$$(2.4) \quad (\mu^0 \delta_1 - \lambda) \cdot [\phi'(\bar{b}_1^*) \bar{b}_1^*] = \min_{b_1^* \in B_1^*} (\mu^0 \delta_1 - \lambda) \cdot [\phi'(\bar{b}_1^*) b_1^*] \\ \text{for some } \mu^0 \geq 0,$$

where $\delta_1 = (1, 0, \dots, 0) \in E_n$.

Remark. Relation (2.2) generalizes the Weierstrass E -condition, relation (2.3) generalizes the transversality conditions at the initial point and describes the dependence on parameters, and relation (2.4) generalizes the transversality conditions at the endpoint.

Theorem 2.2 is of particular interest in the case [1] when $t \rightarrow R^\#(t) \subset R$ is a given mapping from T to the class of nonempty subsets of R , and \mathfrak{R} is the class of measurable relaxed controls σ such that the probability measure $\sigma(t)$ is supported on the closure of $R^\#(t)$ for all $t \in T$. We may then assert [1, Theorem 2.6] that in a large class of problems there exist a relaxed control $\bar{\sigma}$ and a point \bar{b} that yield the restricted minimum assumed in Theorem 2.2; and we may verify a priori the other assumptions of Theorem 2.2. We are then able to state that a minimizing control $\bar{\sigma}$ and point \bar{b} exist and either satisfy condition (2.1) or conditions (2.2), (2.3), and (2.4). Since these relations often admit only a finite number of solutions, we can determine a minimizing $\bar{\sigma}$ and \bar{b} ; in this sense, [1, Theorem 2.6] and Theorem 2.2 often provide constructive conditions for minimum.

3. Functions of controls defined by ordinary differential equations. We shall now illustrate the use of Theorem 2.2 in certain standard problems of the control theory, postponing the proof of the results presented in this section to §5. Let T be the closed interval $[t_0, t_1]$ of the real axis, R a separable metric space, $R^\#$ a mapping from T to the class of nonempty subsets of R , $B_0 \subset E_n$, $B_1 \subset E_n$, and $g: E_n \times T \times R \rightarrow E_n$. In this section, and in §5, the words *measure* and *measurable* will be used in the sense of Lebesgue and $|A|$ will represent the measure of $A \subset T$.

Let \mathfrak{R}' be a class of mappings $\rho: T \rightarrow R$ such that $t \rightarrow g(v, t, \rho(t))$ is measurable on T for every $v \in E_n$ and $\rho \in \mathfrak{R}'$ and

$$[\rho_i, A_i (i = 1, \dots, k); \rho] \in \mathfrak{R}'$$

if k is a positive integer, each A_i is a denumerable union of intervals, and $\rho \in \mathfrak{R}'$, $\rho_i \in \mathfrak{R}'$, $i = 1, \dots, k$. We shall henceforth refer to elements of \mathfrak{R}' as "measurable" mappings (as distinguished from measurable mappings). We set $\mathfrak{R} = \{\rho \in \mathfrak{R}' \mid \rho(t) \in R^\#(t) \text{ on } T\}$.

For $\rho \in \mathfrak{R}$ and $b_0 \in B_0$, we consider an absolutely continuous function $y = y(\cdot; \rho, b_0)$ on T such that

$$(3.1) \quad \frac{dy(t)}{dt} = \dot{y}(t) = g(y(t), t, \rho(t)) \quad \text{a.e. in } T$$

and

$$(3.2) \quad y(t_0) = b_0.$$

We wish to investigate certain properties of a point $\bar{b}_0 \in B_0$ and a mapping $\bar{\rho} \in \mathcal{R}$ that minimize $y^1(t_1; \rho, b_0)$ on $\mathcal{R} \times B_0$ subject to the restriction that $y(t_1; \rho, b_0) \in B_1$.

We shall say that a sequence $\{M_j\}_{j=1}^\infty$ of closed subsets of T is "regular at \bar{t} " if $|M_j| \rightarrow 0$ as $j \rightarrow \infty$, $\bar{t} \in M_j$, and diameter $(M_j) < c |M_j|$ for some positive c and all $j = 1, 2, \dots$. We shall say that a "measurable" mapping $\rho^*: T \rightarrow R$ is "admissible at \bar{t} " if $\rho^*(t) \in R^*(t)$ on T and

$$\lim_{j \rightarrow \infty} (1/|M_j|) \int_{M_j} g(v, t, \rho^*(t)) dt = g(v, \bar{t}, \rho^*(\bar{t}))$$

for all $v \in E_n$ and all sequences $\{M_j\}$ that are regular at \bar{t} .

We set

$$\mathcal{R}^*(\bar{t}) = \{\rho^* | \rho^* \text{ is admissible at } \bar{t}\},$$

$$R^*(\bar{t}) = \{\rho^*(\bar{t}) | \rho^* \in \mathcal{R}^*(\bar{t})\}.$$

We shall also write $\bar{R}^*(\bar{t})$, \bar{M}_j , etc. to represent closures of the sets $R^*(\bar{t})$, M_j , etc.

ASSUMPTION 3.1. For every $v \in E_n$, $r \in R$, and $t \in T$, the function $g(v, t, \cdot)$ is continuous on R , $g(v, \cdot, r)$ is measurable on T , and $g(\cdot, t, r)$ is continuous and has continuous first order derivatives on E_n . Furthermore, for every v in E_n there exists an integrable function ψ_v on T such that $|g(v, t, r)| \leq \psi_v(t)$ on $T \times R$. Finally, for every bounded subset D of E_n there exists an integrable function ψ_D on T such that $|g_v(v, t, r)| \leq \psi_D(t)$ on $D \times T \times R$. Here $|g| = (\sum_{j=1}^n (g^j)^2)^{1/2}$, g_v is the matrix $(\partial g^i / \partial v^j)$, and $|g_v| = \sum_{i,j=1}^n |\partial g^i / \partial v^j|$.

Remark. Assumption 3.1 implies that we may choose as \mathcal{R}' the class of all the measurable functions from T to R .

THEOREM 3.2. Let $(\bar{\rho}, \bar{b}_0)$ minimize $y^1(t_1; \rho, b_0)$ among all points b_0 in B_0 and all "measurable" mappings ρ such that $\rho(t) \in R^*(t)$ on T and $y(t_1; \rho, b_0) \in B_1$, and let Assumption 3.1 be satisfied. Let $\bar{y} = y(\cdot; \bar{\rho}, \bar{b}_0)$, $\bar{b}_1 = \bar{y}(t_1)$, and let, for $k = 0, 1$, B_k^* be a convex set in E_{m_k} and $\phi_k = (\phi_k^1, \dots, \phi_k^n): E_{m_k} \rightarrow E_n$ be a continuously differentiable mapping such that $\phi_k(B_k^*) \subset B_k$ and $\phi_k(\bar{b}_k^*) = \bar{b}_k$ for some $\bar{b}_k^* \in B_k^*$. Let A_k be, for $k = 0, 1$, the matrix $(\partial \phi_k^i / \partial b_k^{*j})$ evaluated at \bar{b}_k^* , and let A_k^i be the i th row of this matrix. Then either

$$(3.3) \quad A_1^1 \cdot \bar{b}_1^* = \min_{b_1^* \in B_1^*} A_1^1 \cdot b_1^*,$$

or there exists an absolutely continuous function $z: T \rightarrow E_n$ such that

$$(3.4) \quad \dot{y}(t) = g(\bar{y}(t), t, \bar{p}(t)) \quad \text{a.e. in } T,$$

$$(3.5) \quad \dot{z}(t) = -g_v^T(\bar{y}(t), t, \bar{p}(t))z(t) \quad \text{a.e. in } T$$

(where g_v^T is the transpose of the matrix g_v),

$$(3.6) \quad |z(t)| \neq 0 \quad \text{on } T,$$

$$(3.7) \quad z(t) \cdot g(\bar{y}(t), t, \bar{p}(t)) = \min_{r \in R^*(t)} z(t) \cdot g(\bar{y}(t), t, r) \quad \text{a.e. in } T,$$

$$(3.8) \quad z(t_0) \cdot A_0 \bar{b}_0^* = \min_{b_0^* \in B_0^*} z(t_0) \cdot A_0 b_0^*,$$

and

$$(3.9) \quad (\gamma \delta_1 - z(t_1)) \cdot A_1 \bar{b}_1^* = \min_{b_1^* \in B_1^*} (\gamma \delta_1 - z(t_1)) \cdot A_1 b_1^*$$

for some $\gamma \geq 0$. Here $\delta_1 = (1, 0, \dots, 0) \in E_n$.

In particular, if $R^*(t) = R$ on T and \mathcal{R}' contains the constant mapping $t \rightarrow r$ for all r in a dense denumerable subset R_∞ of R , then $\bar{R}^*(t)$ can be replaced by R in statement (3.7).

By combining [1, Theorem 3.1] and Theorem 3.2, we can prove the existence of a minimizing relaxed control $\bar{\sigma}$ and a point \bar{b}_0 and can state some of their characteristic properties.

ASSUMPTION 3.3.

- (i) R is compact;
- (ii) B_0 is compact and B_1 is closed;
- (iii) there exists an integrable function ψ on T such that $|g_v(v, t, r)| \leq \psi(t)$ on $E_n \times T \times R$;
- (iv) let $\mathcal{R}^* = \{\rho: T \rightarrow R \mid \rho(t) \in R^*(t) \text{ on } T \text{ and } \rho \text{ is measurable}\}$. Then for every ϵ there exists a closed subset T_ϵ of T , of measure at least $|T| - \epsilon$, with the property that (a) for every $\bar{t} \in T_\epsilon$ and every $r \in R^*(\bar{t})$ there exists a mapping $\rho \in \mathcal{R}^*$, continuous at \bar{t} when restricted to T_ϵ , and such that the distance from $\rho(\bar{t})$ to r is at most ϵ ; and (b) for every $\bar{t} \in T_\epsilon$ and every $h > 0$ there exists a positive $\delta = \delta(h, \bar{t})$ such that $\bar{R}^*(t)$ and $R^*(\bar{t})$ are in the h -neighborhood of each other if $t \in T_\epsilon$ and $|t - \bar{t}| \leq \delta$.

Now let S be the class of regular Borel probability measures on R . It is well known [9, p. 426] that a metric can be defined on S such that S is separable and the convergence in S is the weak convergence of measures: that is, a sequence s_1, s_2, \dots converges to s in S if $\int_R c(r) s_j(dr) \rightarrow \int_R c(r) s(dr)$ as $j \rightarrow \infty$ for every continuous $c: R \rightarrow E_1$. Let S^* be the set of mappings σ from T to S such that $\sigma(\bar{R}^*(t); t) = 1$ on T and

$t \rightarrow \int_R c(r)\sigma(dr; t)$ is measurable on T for every continuous $c: R \rightarrow E_1$.

Here $\sigma(R'; t)$ is the $\sigma(t)$ -measure of a subset $R' \subset R$.

We refer to an absolutely continuous function $\xi: T \rightarrow E_n$ as a "relaxed curve" if $\xi(t_0) \in B_0$ and $\xi(t)$ belongs, a.e. in T , to the convex closure of the set $\{g(\xi(t), t, r) \mid r \in R^\#(t)\}$. This definition is equivalent, in view of our assumptions and of [1, Theorem 3.1], to the statement that $\xi = \xi(\cdot; \sigma, b_0)$ satisfies the relations

$$\dot{\xi}(t) = \int_R g(\xi(t), t, r)\sigma(dr; t) \quad \text{a.e. in } T,$$

$$\xi(t_0) = b_0,$$

for some $\sigma \in S^*$ and $b_0 \in B_0$. (This definition is also consistent with the one in [1, §3] for $A = E_n$.)

THEOREM 3.4. *Let B_0, B_1, T, R, R^* , and g satisfy Assumptions 3.1 and 3.3 and assume that $\xi(t_1; \sigma', b_0') \in B_1$ for some $\sigma' \in S^*$ and $b_0' \in B_0$. Then there exist a relaxed control $\bar{\sigma}$ and a point $\bar{b}_0 \in B_0$ that minimize $\xi^1(t_1; \sigma, b_0)$ on $S^* \times B_0$ subject to the condition that $\xi(t_1; \sigma, b_0) \in B_1$; and the corresponding minimizing relaxed curve $\bar{\xi} = \xi(\cdot; \bar{\sigma}, \bar{b}_0)$ can be uniformly approximated on T by a sequence ξ_1, ξ_2, \dots of absolutely continuous curves such that $\xi_j(t) = g(\xi_j(t), t, \rho_j(t))$ a.e. in $T, j = 1, 2, \dots$, the mappings ρ_j are measurable, and $\rho_j(t) \in R^*(t)$ on T .*

Let $f(v, t, s) = \int_R g(v, t, r)s(dr)$ on $E_n \times T \times S$, let $S^*(t) = \{s \in S \mid s(\bar{R}^*(t)) = 1\}$ for $t \in T$, let $\bar{b}_1 = \xi(t_1; \bar{\sigma}, \bar{b}_0)$, and let E_{m_k}, B_k^*, ϕ_k , and A_k be defined as in the statement of Theorem 3.2. Then either condition (3.3) or conditions (3.4) through (3.9) of Theorem 3.2 are satisfied, with $\bar{y}, g, \bar{\rho}, R^*$, and r replaced by, respectively, $\bar{\xi}, f, \bar{\sigma}, S^*$, and s .

Furthermore, condition (3.7) of Theorem 3.2 implies that, a.e. in T ,

$$(3.10) \quad z(t) \cdot g(\bar{\xi}(t), t, \bar{r}) = \min_{r \in R} z(t) \cdot g(\bar{\xi}(t), t, r)$$

for every \bar{r} in the support of $\bar{\sigma}(t)$, if $R^*(t) = R$ on T .

4. Proof of Theorem 2.2. The proof of Theorem 2.2 is essentially contained in the lemma that follows and that resembles, in many respects, Lemma 3.1 of [3, p. 132]. The convex set W is patterned after a construction of McShane [6, pp. 17–18]. Brouwer's fixed-point theorem appears to have been first applied in a similar context by H. Halkin [7, p. 75].

LEMMA 4.1. *Let $(\bar{\rho}, \bar{b})$ minimize $x^1(\rho, b)$ in $\mathcal{R} \times B$ subject to the conditions $x^l(\rho, b) = 0, l = 2, \dots, n$. Let $(T^*, \mathcal{R}^*, \mathcal{N})$ define local variations for x in $\mathcal{R} \times B$ at $(\bar{\rho}, \bar{b})$. Then there exists a nonvanishing vector λ in E_n such that $\lambda^1 \geq 0$,*

$$(4.1) \quad \lambda \cdot Dx(\bar{p}, \bar{b}; t^*, r) \geq 0 \quad \text{for all } t^* \in T^* \quad \text{and } r \in R^*(t^*),$$

and

$$(4.2) \quad \lambda \cdot Dx(\bar{p}, \bar{b}; b) \geq 0 \quad \text{for all } b \in B.$$

Proof. We shall use the notation that we have introduced in §2. Let $V_1 = \{Dx(\bar{p}, \bar{b}; t^*, r) \mid t^* \in T^*, r \in R^*(t^*)\}$, $V_2 = \{Dx(\bar{p}, \bar{b}; b) \mid b \in B\}$, and let W be the convex cone in E_n generated by $V_1 \cup V_2$; that is, $W = \{a^1 v_1 + \cdots + a^n v_n \mid v_i \in V_1 \cup V_2, a^i \geq 0, i = 1, \dots, n\}$. Assume now, by way of contradiction, that there exists no vector λ with the stated properties. Then we can easily deduce from elementary properties of convex sets that there exists a point $w = (w^1, 0, \dots, 0)$ in the interior of W , linearly independent vectors (points) $w_i \in W$, and positive numbers c^i , $i = 1, \dots, n$, such that

$$(4.3) \quad w^1 < 0 \quad \text{and} \quad w = \sum_{i=1}^n c^i w_i.$$

By the definition of W , there exist points $t^{ij} \in T^*$, controls $\rho^{ij} \in \mathcal{R}^*(t^{ij})$, points $b^{ij} \in B$ and numbers $a^{ij} \geq 0$, $i, j = 1, \dots, n$, such that

$$(4.4) \quad w_i = \sum_{j=1}^n a^{ij} \cdot Dx(\bar{p}, \bar{b}; t^{ij}, \rho^{ij}, b^{ij}), \quad i = 1, \dots, n,$$

where, for each i, j , $Dx(\bar{p}, \bar{b}; t^{ij}, \rho^{ij}, b^{ij})$ either represents $Dx(\bar{p}, \bar{b}; t^{ij}, \rho^{ij})$ (and is independent of b^{ij}), or represents $Dx(\bar{p}, \bar{b}; b^{ij})$ (and is independent of t^{ij} and ρ^{ij}). The matrix $(w_i^l, i, l = 1, \dots, n)$, is nonsingular since the vectors w_i are linearly independent.

Let now $\bar{\alpha}$ be sufficiently small so that the sets $N_{k_1}(t^{i_1 j_1}, \alpha)$ and $N_{k_2}(t^{i_2 j_2}, \beta)$ are disjoint if $(k_1, t^{i_1 j_1}) \neq (k_2, t^{i_2 j_2})$, $\alpha \leq \bar{\alpha}$ and $\beta \leq \bar{\alpha}$, let $\bar{\alpha} < 1$, and let

$$\Delta = \left\{ \delta \in E_n \mid 0 \leq \delta^i \leq \bar{\alpha} / \left(\sum_{k,j=1}^n a^{kj} \right), \quad i = 1, \dots, n \right\}.$$

For every $\delta \in \Delta$, let $\omega^{ij}(\delta) = a^{ij} \delta^i$ and $\theta^{ij}(\delta) = 0$ (respectively, $\omega^{ij}(\delta) = 0$ and $\theta^{ij}(\delta) = a^{ij} \delta^i$) if $Dx(\bar{p}, \bar{b}; t^{ij}, \rho^{ij}, b^{ij})$ represents $Dx(\bar{p}, \bar{b}; t^{ij}, \rho^{ij})$ (respectively, $Dx(\bar{p}, \bar{b}; b^{ij})$). We observe that the sets $N_{n_j+i-n}(t^{ij}, \omega^{ij}(\delta))$ are disjoint and $\sum_{i,j=1}^n \theta^{ij}(\delta) \leq 1$ for $\delta \in \Delta$.

We now consider, for each $\delta \in \Delta$, the "perturbed" mapping $\rho'(\delta) = \rho'(t^\square, \rho^\square, \omega^\square(\delta)) = [\rho^{ij}, N_{n_j-n+i}(t^{ij}, \omega^{ij}(\delta)) \mid (i, j = 1, \dots, n); \bar{p}]$ in \mathcal{R} and the "perturbed" point $b'(\delta) = \theta^\square(\delta) \circ b^\square$ in B . By condition (b) of Definition 2.1, the function $\delta \rightarrow \xi(\delta) = \xi(\omega^\square(\delta), \theta^\square(\delta)) = x(\rho'(\delta), b'(\delta))$ from Δ to E_n is continuous in some neighborhood Δ' (relative to Δ) of the origin 0 of E_n and has a differential at 0 (relative to Δ). Furthermore, by (4.4), the right-hand derivative

$$\begin{aligned}
& \left. \frac{\partial \tilde{\xi}(\delta)}{\partial \delta^k} \right|_{\delta=0} \\
&= \sum_{i,j=1}^n \left[\frac{\partial \xi(\omega^{\square}(\delta), \theta^{\square}(\delta))}{\partial \omega^{ij}} \cdot \frac{\partial \omega^{ij}(\delta)}{\partial \delta^k} + \frac{\partial \xi(\omega^{\square}(\delta), \theta^{\square}(\delta))}{\partial \theta^{ij}} \cdot \frac{\partial \theta^{ij}(\delta)}{\partial \delta^k} \right] \Big|_{\delta=0} \\
&= \sum_{j=1}^n D x(\bar{\rho}, \bar{b}; t^{kj}, \rho^{kj}, b^{kj}) a^{kj} = w_k, \quad k = 1, \dots, n.
\end{aligned}$$

Thus the Jacobian matrix $\tilde{\xi}_\delta(0) = (\partial \tilde{\xi}^i(0)/\partial \delta^j)_{i,j=1}^n$ is nonsingular.

Let $a(\delta) = (\tilde{\xi}_\delta(0))^{-1}(\tilde{\xi}(\delta) - \tilde{\xi}(0) - \tilde{\xi}_\delta(0)\delta)$, and let $c = (c^1, \dots, c^n)$, where the c^i , $i = 1, \dots, n$, are as in (4.3). We shall now show that the equation

$$(4.5) \quad \delta = \gamma c - a(\delta)$$

has a solution $\delta(\gamma)$ for all sufficiently small positive γ . Indeed, since $\tilde{\xi}(\cdot)$ has a differential at 0 (relative to Δ), there exists a positive β_0 such that $|a^i(\delta)| \leq \frac{1}{4}c_{\min}\delta_{\max}/c_{\max}$ and $\delta \in \Delta'$ if $0 \leq \delta^i \leq \beta_0$, $i = 1, \dots, n$, where $c_{\min} = \min_i c^i$, $c_{\max} = \max_i c^i$, and $\delta_{\max} = \max_i \delta^i$. Let $0 < \beta \leq \beta_0$, $\gamma = \frac{1}{2}\beta/c_{\max}$, and $\Delta_\gamma'' = \{\delta \in \Delta' \mid |\delta^i - \gamma c^i| \leq \frac{1}{2}\gamma c_{\min}, i = 1, 2, \dots, n\}$. Then we can easily verify that Δ_γ'' is homeomorphic to a closed ball in E_n , and $\gamma c - a(\delta)$ is a continuous mapping of Δ_γ'' into itself. Thus, by Brouwer's fixed-point theorem, there exists $\delta = \delta(\gamma)$ satisfying equation (4.5). It follows then from relations (4.3) and (4.5) that

$$\tilde{\xi}(\delta(\gamma)) - \tilde{\xi}(0) = \gamma \tilde{\xi}_\delta(0)c = \gamma \sum_{i=1}^n c^i w_i = \gamma w = (\gamma w^1, 0, \dots, 0);$$

hence,

$$\begin{aligned}
\tilde{\xi}^1(\delta(\gamma)) &= \tilde{\xi}^1(0) + \gamma w^1 < \tilde{\xi}^1(0) = x^1(\bar{\rho}, \bar{b}), \\
\tilde{\xi}^l(\delta(\gamma)) &= 0, \quad l = 2, \dots, n.
\end{aligned}$$

Since $\rho'(\delta(\gamma)) \in \mathcal{R}$ and $b'(\delta(\gamma)) \in B$ for all $\beta \leq \beta_0$ and $\delta \in \Delta_\gamma'' \subset \Delta$, and since $\delta(\gamma) \in \Delta_\gamma''$ and $\tilde{\xi}(\delta(\gamma)) = x(\rho'(\delta(\gamma)), b'(\delta(\gamma)))$, we conclude that, contrary to assumption, $(\bar{\rho}, \bar{b})$ does not minimize $x^1(\rho, b)$ subject to the restriction that $x^l(\rho, b) = 0$, $l = 2, \dots, n$. This completes the proof of the lemma.

Proof of Theorem 2.2. Let $c = (b, b_1^*)$ for $b \in B$ and $b_1^* \in B_1^*$, let $\bar{c} = (\bar{b}, \bar{b}_1^*)$, and let $C = B \times B_1^*$. Then $(\bar{\rho}, \bar{c})$ minimizes $x^1(\rho, b)$ on $\mathcal{R} \times C$ subject to the restrictions $x^l(\rho, b) - \phi^l(b_1^*) = 0$, $l = 1, 2, \dots, n$.

Let the function $y = (y^0, y^1, \dots, y^n)$ on $\mathcal{R} \times C$ be defined by

$$\begin{aligned}
y^0(\rho, c) &= y^0(\rho, b, b_1^*) = x^1(\rho, b), \\
y^l(\rho, c) &= y^l(\rho, b, b_1^*) = x^l(\rho, b) - \phi^l(b_1^*), \quad l = 1, \dots, n.
\end{aligned}$$

Then we verify that $(T^*, \mathfrak{R}^*, \mathfrak{N})$ defines local variations for y in $\mathfrak{R} \times C$ at (\bar{p}, \bar{c}) . It follows then, by Lemma 4.1, that there exists a nonvanishing vector $\mu = (\mu^0, \mu^1, \dots, \mu^n)$ in E_{n+1} and a vector $\lambda = (\mu^0 + \mu^1, \mu^2, \dots, \mu^n)$ in E_n such that $\mu^0 \geq 0$ and

$$(4.6) \quad \mu \cdot Dy(\bar{p}, \bar{c}; t^*, r) = \lambda \cdot Dx(\bar{p}, \bar{b}; t^*, r) \geq 0$$

for all $t^* \in T^*$ and $r \in R^*(t^*)$, and

$$(4.7) \quad \begin{aligned} \mu \cdot Dy(\bar{p}, \bar{c}; c) &= \mu^0 Dx^1(\bar{p}, \bar{b}; b) + (\lambda - \mu^0 \delta_1) \cdot (Dx(\bar{p}, \bar{b}; b) \\ &\quad - \phi'(\bar{b}_1^*)(b_1^* - \bar{b}_1^*)) \geq 0 \end{aligned}$$

for all $b \in B$ and $b_1^* \in B_1^*$, where $\delta_1 = (1, 0, \dots, 0) \in E_n$. We observe that $Dx(\bar{p}, \bar{b}; \bar{b}) = 0$; hence, setting $b = \bar{b}$ in (4.7), it follows that

$$(\lambda - \mu^0 \delta_1) \cdot \phi'(\bar{b}_1^*)(b_1^* - \bar{b}_1^*) \leq 0 \quad \text{for all } b_1^* \in B_1^*.$$

Since μ is nonvanishing, either λ is nonvanishing or $\mu = (\mu^0, -\mu^0, 0, \dots, 0)$, $\mu^0 > 0$, and

$$\phi^{1'}(\bar{b}_1^*)\bar{b}_1^* = \min_{b_1^* \in B_1^*} \phi^{1'}(\bar{b}_1^*)b_1^*.$$

5. Functions of controls defined by ordinary differential equations.

Proofs. We shall use the notation of §3, and we shall make, at first, the same assumptions as in Theorem 3.2.

For any integrable function f from T to some Euclidean space, let $T'(f)$ be the set of all the points t^* in T such that $|f(t^*)|$ is finite and

$$\lim_{j \rightarrow \infty} \frac{1}{|M_j|} \int_{M_j} f(t) dt = f(t^*)$$

for all sequences $\{M_j\}_{j=1}^\infty$ of closed subsets of T that are regular at t^* . It is well known [8, Theorem (6.3), p. 118] that the set $T'(f)$ has measure $|T|$.

Now let D^0 be a bounded convex open set containing the range of \bar{y} , let $\psi_0 = \psi_{D^0}$ be defined as in Assumption 3.1, let D_∞ be a dense denumerable subset of D^0 and R_∞ a dense denumerable subset of $\bigcup_{t \in T} R^*(t)$, and let $g(v, \cdot, \bar{p}(\cdot))$ be the function $t \rightarrow g(v, t, \bar{p}(t))$. Then

$$T^* = \bigcap_{v \in D_\infty, r \in R_\infty} [T'(g(v, \cdot, r)) \cap T'(g(v, \cdot, \bar{p}(\cdot)))] \cap T'(\psi_0) \cap [t_0, t_1]$$

has measure $|T|$. We also verify that $T'(g(v, \cdot, r)) \supset T^*$ and $T'(g(v, \cdot, \bar{p}(\cdot))) \supset T^*$ for all $v \in D^0$ and $r \subset R_\infty$. Indeed, let $t^* \in T^*$, $v \in D^0$, $r \in R_\infty$, and let v_1, v_2, \dots be a sequence in D_∞ converging to v . Then, by Assumption 3.1,

$$\limsup_{j \rightarrow \infty} \left| \frac{1}{|M_j|} \int_{M_j} g(v, t, r) dt - g(v_i, t^*, r) \right|$$

$$\begin{aligned} &\leq \limsup_{j \rightarrow \infty} \frac{1}{|M_j|} \int_{M_j} |g(v, t, r) - g(v_i, t, r)| dt \\ &\leq \limsup_{j \rightarrow \infty} |v - v_i| \frac{1}{|M_j|} \int_{M_j} \psi_0(t) dt = \psi_0(t^*) |v - v_i|, \quad i = 1, 2, \dots, \end{aligned}$$

for every sequence $\{M_j\}$ that is regular at t^* . Since $g(\cdot, t^*, r)$ is continuous on D^0 , we conclude that $T'(g(v, \cdot, r)) \supset T^*$. We similarly show that $T'(g(v, \cdot, \bar{p}(\cdot))) \supset T^*$ for all $v \in D^0$.

We next define sets $N_k(t, \alpha)$ and the corresponding collection \mathfrak{N} . Let $m = n^2$, and let

$$\begin{aligned} \eta_{k-1, i} &= 2^{-mi+m-k+1}, \\ N_k &= \bigcup_{i=1}^{\infty} (\eta_{k, i}, \eta_{k-1, i}], \end{aligned}$$

and

$$N_k(t) = (t + N_k) \cap T, \quad t \in T, \quad k = 1, \dots, m, \quad i = 1, 2, \dots.$$

We then define $N_k(t, \alpha)$ for $\alpha \geq 0$ as $N_k(t) \cap (t, t + \beta_k(t, \alpha)]$, where $\beta_k(t, \alpha)$ is nonnegative and such that $|N_k(t, \alpha)| = \min(\alpha, |N_k(t)|)$. We observe that $\text{diameter}(N_k(t, \alpha)) \leq 2^m |N_k(t, \alpha)|$ for all t and α , and $|N_k(t, \alpha)| = \alpha$ for sufficiently small α and $t < t_1$.

We shall henceforth use the above definitions of T^* and \mathfrak{N} , as well as the definitions of \mathcal{R}^* and R^* of §3. We shall also use the notation of Definition 2.1. Let $t^{ij} \in T^*$ and $\rho^{ij} \in \mathcal{R}^*(t^{ij})$, $i, j = 1, \dots, n$, and let $\rho' = \rho'(t^\square, \rho^\square, \omega^\square) = [\rho^{ij}, N_{nj-n+i}(t^{ij}, \omega^{ij})(i, j = 1, \dots, n); \bar{\rho}]$ for all $\omega^\square \in \Omega(t^\square)$. Finally, let $B = B_0^*$, $\bar{b} = \bar{b}_0^*$, and for $\rho \in \mathcal{R}$ and $b \in B$, let the absolutely continuous function $y = y(\cdot; \rho, b): T \rightarrow E_n$ be the solution of the system

$$\begin{aligned} \dot{y}(t) &= g(y(t), t, \rho(t)) \quad \text{a.e. in } T, \\ y(t_0) &= \phi_0(b). \end{aligned}$$

It follows from Assumption 3.1 and from well-known theorems that there exists a neighborhood \tilde{B} of \bar{b} in B such that the function y , as just defined, exists, is unique, has its range contained in D^0 , and depends continuously on b , uniformly in ρ , for all $b \in \tilde{B}$ and all "measurable" ρ such that the set $\{t \in T \mid \rho(t) \neq \bar{\rho}(t)\}$ has a sufficiently small measure.

LEMMA 5.1. Let $\rho' = \rho'(t^\square, \rho^\square, \omega^\square)$. For all $t \in T$, $b \in \tilde{B}$, $t^{ij} \in T^*$, and $\rho^{ij} \in \mathcal{R}^*(t^{ij})$, $i, j = 1, \dots, n$, the function $(\omega^\square, b) \rightarrow y(t; \rho', b)$ is continuous in some neighborhood Γ of $(0^\square, \bar{b})$ in $\Omega(t^\square) \times \tilde{B}$; and, for all $i, j = 1, \dots, n$, the limit defining the right-hand derivative of $y(t; \rho', b)$ with respect to ω^{ij} at 0 exists and is uniform in Γ , and this derivative is a continuous function of (ω^\square, b) in Γ . Similarly,

$$\lim_{\theta \rightarrow +0} \frac{1}{\theta} (y(t; \rho', b_1 + \theta(b_2 - b_1)) - y(t; \rho', b_1))$$

defining the right-hand derivative $\partial y(t; \rho', b_1 + \theta(b_2 - b_1))/\partial \theta|_{\theta=0}$ exists, this limit is uniform in $\Gamma \times \Gamma$, and it is continuous in $\Gamma \times \Gamma$.

Finally, let $x(\rho, b) = y(t_1; \rho, b)$. Then

$$D_{\mathfrak{N}_k} x(\bar{\rho}, \bar{b}; t^*, \rho^*) = Dx(\bar{\rho}, \bar{b}; t^*, \rho^*(t^*)) = Z(t^*)(g(y(t^*), t^*, \rho^*(t^*)) - g(y(t^*), t^*, \bar{\rho}(t^*))), \quad k = 1, 2, \dots, n^2, \quad t^* \in T^*, \quad \rho^* \in \mathfrak{R}^*(t^*),$$

and

$$Dx(\bar{\rho}, \bar{b}; b) = Z(t_0) \frac{\partial \phi_0(\bar{b})}{\partial b} (b - \bar{b}), \quad b \in B,$$

where the matrix function Z is the solution of the system $\dot{Z}(t) = -Z(t)g_v(\bar{y}(t), t, \bar{\rho}(t))$ a.e. in T , $Z(t_1) = I$ (the unit matrix).

Proof. Let $t^{\bar{i}\bar{j}}$ and $\rho^{\bar{i}\bar{j}}$ be fixed. For fixed \bar{i} and \bar{j} in $\{1, 2, \dots, n\}$, let $t^* = t^{\bar{i}\bar{j}}$, $\rho^* = \rho^{\bar{i}\bar{j}}$, and $M(\alpha) = N_{n\bar{j}-n+\bar{i}}(t^*, \alpha)$ for $\alpha \geq 0$.

We observe that, for every sequence $\alpha_1, \alpha_2, \dots$ converging to $+0$, the sequence $\{\bar{M}_a\} = \{\bar{M}(\alpha_a)\}_{a=1}^\infty$ is regular at t^* , $|\bar{M}_a - M_a| = 0$ for all a , and $|M(\alpha)| = \alpha$ for sufficiently small α . It follows that, for all $v \in D^0$,

$$\lim_{\alpha \rightarrow +0} \frac{1}{\alpha} \int_{M(\alpha)} g(v, t, \rho(t)) dt = g(v, t^*, \rho(t^*))$$

if $\rho = \bar{\rho}$, $\rho = \rho^*$, or $\rho(t) = r \in R_\infty$ on T .

We next consider $y(t; \rho', b)$ as a function of (ω^\square, b) . The measure of the set $\{t \in T \mid \rho'(t^\square, \rho^\square, \bar{\omega}^\square) \neq \rho'(t^\square, \rho^\square, \omega^\square)\}$ converges to 0 uniformly in $\Omega(t^\square)$ when $|\omega^\square - \bar{\omega}^\square| = \sum_{i,j=1}^n |\omega^{ij} - \bar{\omega}^{ij}| \rightarrow 0$. Furthermore, $|g(v, \cdot, r)|$ and $|g_v(v, \cdot, r)|$ are bounded by some integrable function ψ_1 on $D^0 \times R$. We conclude, using standard arguments, that $y(t; \rho', b)$ is a uniformly continuous function of (ω^\square, b, t) and $y(t; \rho', b) \in D^0$ in $\Gamma \times T$, where Γ is some neighborhood of $(0^\square, \bar{b})$ in $\Omega(t^\square) \times B$.

Now we fix b and sufficiently small $\omega^{ij} ((i, j) \neq (\bar{i}, \bar{j}))$ as well as $\bar{i}, \bar{j}, t^\square$, and ρ^\square , and set $\bar{\rho}(\alpha) = \rho'(t^\square, \rho^\square, \omega^\square)$ and $\bar{y}(t; \alpha) = y(t; \bar{\rho}(\alpha), b)$ for $\omega^{ij} = \alpha \geq 0$ and $t \in T$. Then, for sufficiently small α , $\bar{\rho}(t; \alpha) = \bar{\rho}(t; 0)$ for $t \in T - M(\alpha)$, $\bar{\rho}(t; \alpha) = \rho^*(t)$ and $\bar{\rho}(t; 0) = \bar{\rho}(t)$ for $t \in M(\alpha)$, $\bar{y}(t; \alpha) = \bar{y}(t; 0)$ for $t \leq t^*$, and, for $t > t^*$,

$$\begin{aligned} \Delta(t; \alpha) &= \frac{1}{\alpha} (\bar{y}(t; \alpha) - \bar{y}(t; 0)) \\ &= \frac{1}{\alpha} \int_{t^*}^t (g(\bar{y}(\theta; \alpha), \theta, \bar{\rho}(\theta; \alpha)) - g(\bar{y}(\theta; 0), \theta, \bar{\rho}(\theta; 0))) d\theta; \end{aligned}$$

hence,

$$\begin{aligned}
 \Delta(t; \alpha) &= \int_{[t^*, t] - M(\alpha)} g_v(\tilde{y}(\theta; 0), \theta, \tilde{p}(\theta; 0)) \Delta(\theta; \alpha) d\theta \\
 (5.1) \quad &+ \frac{1}{\alpha} \int_{M(\alpha)} (g(\tilde{y}(\theta; \alpha), \theta, \rho^*(\theta)) - g(\tilde{y}(\theta; 0), \theta, \bar{p}(\theta))) d\theta \\
 &+ \int_{[t^*, t] - M(\alpha)} (g_v(\tilde{y}(\theta; \alpha), \theta, \tilde{p}(\theta; 0)) \\
 &\quad - g_v(\tilde{y}(\theta; 0), \theta, \tilde{p}(\theta; 0))) \Delta(\theta; \alpha) d\theta,
 \end{aligned}$$

where $\tilde{y}(\theta; \alpha)$ is, for each θ and α , intermediate between $\tilde{y}(\theta; \alpha)$ and $\tilde{y}(\theta; 0)$. Since $\tilde{y}(\cdot; \alpha)$ converges uniformly to $\tilde{y}(\cdot; 0)$ as $\alpha \rightarrow 0$, $|g_v(v, t, r)| \leq \psi_0(t)$ on $D^0 \times T \times R$, $\rho^* \in \mathcal{R}^*(t^*)$, and $T^* \subset T'(\psi_0)$, we can assert that

$$\begin{aligned}
 \lim_{\alpha \rightarrow +0} \frac{1}{\alpha} \int_{M(\alpha)} (g(\tilde{y}(\theta; \alpha), \theta, \rho^*(\theta)) - g(\tilde{y}(\theta; 0), \theta, \bar{p}(\theta))) d\theta \\
 = \lim_{\alpha \rightarrow +0} \frac{1}{\alpha} \int_{M(\alpha)} (g(\tilde{y}(t^*; 0), \theta, \rho^*(\theta)) - g(\tilde{y}(t^*; 0), \theta, \bar{p}(\theta))) d\theta \\
 = g(\tilde{y}(t^*, 0), t^*, \rho^*(t^*)) - g(\tilde{y}(t^*, 0), t^*, \bar{p}(t^*)),
 \end{aligned}$$

and that this limit is uniform in $\Gamma \times T$. Furthermore,

$$|g_v(\tilde{y}(\theta; \alpha), \theta, \tilde{p}(\theta; 0)) - g_v(\tilde{y}(\theta; 0), \theta, \tilde{p}(\theta; 0))|$$

converges to 0 with α , for each fixed θ in T , uniformly in Γ , because $g_v(\cdot, t, r)$ is continuous, hence uniformly continuous, in some compact set containing D^0 , for every t and r . Moreover, the uniform convergence, hence also the boundedness, of the second term on the right of (5.1) implies that $\Delta(\cdot; \cdot)$ is bounded. Since $|g_v(v, t, r)| \leq \psi_0(t)$ on $D^0 \times T \times R$, it follows then that the last term on the right of (5.1) converges to 0 with α , uniformly in $\Gamma \times T$.

Let $\eta(t) = \lim_{\alpha \rightarrow +0} \Delta(t, \alpha)$ for $t \in T$. We can now conclude that η exists, is unique, that this limit is uniform in $\Gamma \times T$, and that

$$\begin{aligned}
 (5.2) \quad \eta(t) &= \int_{t^*}^t g_v(\tilde{y}(\theta; 0), \theta, \tilde{p}(\theta; 0)) \eta(\theta) d\theta \\
 &+ g(\tilde{y}(t^*, 0), t^*, \rho^*(t^*)) - g(\tilde{y}(t^*, 0), t^*, \bar{p}(t^*)).
 \end{aligned}$$

Now we must investigate the dependence of η on (ω_1^\square, b) . Let (ω_1^\square, b_1) and (ω_2^\square, b_2) be both in Γ and be such that $\omega_1^{ij} = \omega_2^{ij} = 0$, and let $y_1(\cdot)$, $\rho_1(\cdot)$, $\eta_1(\cdot)$ and $y_2(\cdot)$, $\rho_2(\cdot)$, $\eta_2(\cdot)$ represent the corresponding determina-

tions of $\tilde{y}(\cdot; 0)$, $\tilde{\rho}(\cdot; 0)$, and $\eta(\cdot)$. Let also $M = \{t \in T \mid \rho_1(t) \neq \rho_2(t)\}$ and $\Delta(t) = |\eta_1(t) - \eta_2(t)|$. Then (5.2) yields

$$\begin{aligned} \Delta(t) \leq & \int_M \psi_0(\theta) (|\eta_1(\theta)| + |\eta_2(\theta)|) d\theta + \int_{t^*}^t \psi_0(\theta) \Delta(\theta) d\theta \\ & + \int_T |g_v(y_1(\theta), \theta, \rho_1(\theta)) - g_v(y_2(\theta), \theta, \rho_1(\theta))| \cdot |\eta_2(\theta)| d\theta \\ & + |g(y_1(t^*), t^*, \rho^*(t^*)) - g(y_2(t^*), t^*, \rho^*(t^*))| \\ & + |g(y_1(t^*), t^*, \bar{\rho}(t^*)) - g(y_2(t^*), t^*, \bar{\rho}(t^*))|. \end{aligned}$$

We can directly verify from (5.2) that η is uniformly bounded on $\Gamma \times T$. We can show, therefore, as in a previous argument, that the third integral in the last relation converges uniformly (on Γ) to 0 with $|\omega_1^\square - \omega_2^\square| + |b_1 - b_2|$. The first integral converges uniformly to 0 because $|M| \rightarrow 0$ uniformly, and the nonintegrated terms converge to 0 uniformly with $|y_1(t^*) - y_2(t^*)|$. It follows that $\Delta(\cdot) \rightarrow 0$ uniformly on Γ as $|\omega_1^\square - \omega_2^\square| + |b_1 - b_2| \rightarrow 0$.

We can solve equation (5.2), specifically when $\omega^\square = 0^\square$, $b = \bar{b}$, and $r = \rho^*(t^*)$, and find that, for $k = n\bar{j} - n + \bar{i}$,

$$D_{\mathfrak{X}_k} x(\bar{\rho}, \bar{b}; t^*, \rho^*) = \eta(t_1) = Z(t^*)(g(\bar{y}(t^*), t^*, r) - g(\bar{y}(t^*), t^*, \bar{\rho}(t^*))).$$

Thus $D_{\mathfrak{X}_k} x$ is the same for all k and all $\rho^* \in \mathcal{R}^*(t^*)$ such that $\rho^*(t^*) = r$.

Similar arguments prove our assertions concerning $y(t; \rho', b_1 + \theta(b_2 - b_1))$ as a function of θ , and yield the representation of $Dx(\bar{\rho}, \bar{b}; b)$.

This completes the proof of the lemma.

5.1. Completion of the proof of Theorem 3.2. We shall now show that $(T^*, \mathcal{R}^*, \mathfrak{X})$ defines local variations for x in $\mathcal{R} \times B$ at $(\bar{\rho}, \bar{b})$. It is clear that, by construction, the collection \mathfrak{X} satisfies condition (i) of Definition 2.1. Since the sets $N_k(t, \alpha)$ are unions of intervals and $\rho^{ij} \in \mathcal{R}$, $i, j = 1, \dots, n$, the mapping ρ' belongs to \mathcal{R} .

It follows from Lemma 5.1 that the function $(\omega^\square, \theta^\square) \rightarrow \xi(\omega^\square, \theta^\square)$ satisfies condition (b) of Definition 2.1. Indeed, we have shown there that the right-hand partial derivatives of ξ with respect to each ω^{ij} at $\omega^{ij} = 0$ and with respect to each θ^{ij} at $\theta^{ij} = 0$ exist, are continuous, and the limits defining them are uniform for ω^\square and θ^\square sufficiently close to 0^\square . Finally, statement (c) of Definition 2.1 follows directly from Lemma 5.1.

Thus $(T^*, \mathcal{R}^*, \mathfrak{X})$ satisfies the conditions of Definition 2.1, and $Dx(\bar{\rho}, \bar{b}; t^*, r)$ and $Dx(\bar{\rho}, \bar{b}; b)$ have the representations described in Lemma 5.1. All the statements of Theorem 3.2, except statement (3.7), now follow directly from Theorem 2.2 after we set $z(t) = Z^T(t)\lambda$ on T . Fur-

thermore, statement (2.2) implies (3.7), with $\bar{R}^*(t)$ replaced by $R^*(t)$. Since, however, $g(v, t, \cdot)$ is continuous on R for all v and t , we conclude that statement (3.7) is satisfied.

Finally, consider the special case when $R^*(t) = R$ on T and \mathcal{R}' contains all the constant mappings into R_∞ . In that case, for each $r \in R_\infty$ and $t^* \in T^*$, the set $\mathcal{R}^*(t^*)$ contains the constant mapping from T to r , and $\bar{R}^*(t^*) = \bar{R}_\infty = R$.

This completes the proof of Theorem 3.2.

5.2. Proof of Theorem 3.4. The first part of the theorem, concerning the existence of \bar{b}_0 and $\bar{\sigma}$ as well as of the approximating sequences, follows directly from [1, Theorem 3.1]. Next we observe that, for $S^*(t) = \{s \in S \mid s(\bar{R}^*(t)) = 1\}$ on T , S , S^* , f , and s^* satisfy the assumptions made in Theorem 3.2 about R , R^* , g and \mathcal{R} , respectively. Furthermore, since $f(v, t, \tilde{s}) = g(v, t, \tilde{r})$ on $E_n \times T$ for every measure $\tilde{s} = s_r$ concentrated at the single point \tilde{r} , it follows that the set of σ in S^* that are admissible (with respect to f and S^*) at t^* contains $\mathcal{R}^*(t^*)$. Finally, there exists a dense denumerable subset of S containing $\{s_r \mid r \in R_\infty\}$. We may now apply Theorem 3.2, with S , S^* , f , and s^* replacing R , R^* , g , and \mathcal{R} , respectively, and derive directly the second part of Theorem 3.4.

REFERENCES

- [1] J. WARGA, *Functions of relaxed controls*, this Journal, 5 (1967), pp. 628-641.
- [2] ———, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111-128.
- [3] ———, *Necessary conditions for minimum in relaxed variational problems*, Ibid., 4 (1962), pp. 129-145.
- [4] L. W. NEUSTADT, *An abstract variational theory with applications to a broad class of optimization problems. I. General theory*, this Journal, 4 (1966), pp. 505-527.
- [5] ———, *An abstract variational theory with applications to a broad class of optimization problems. II. Applications*, this Journal, 5 (1967), pp. 90-137.
- [6] E. J. MCSHANE, *Necessary conditions in generalized-curve problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 1-27.
- [7] H. HALKIN, *On the necessary condition for optimal control of nonlinear systems*, J. Analyse Math., 12 (1964), pp. 1-82.
- [8] S. SAKS, *Theory of the Integral*, Monografie Matematyczne, 2nd ed., 1937, reprinted by Hafner Publishing Co., New York.
- [9] N. DUNFORD AND J. SCHWARZ, *Linear Operators I*, Interscience, New York, 1964.